Learning the fundamental mid-infrared spectral components of galaxies with non-negative matrix factorization

Article (Published Version)

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **437,** 241–261 (2014) doi:10.1093/mnras/stt1875
Advance Access publication 2013 November 11

Learning the fundamental mid-infrared spectral components of galaxies with non-negative matrix factorization

P. D. Hurley,[1]⋆ S. Oliver,[1] D. Farrah,[2] V. Lebouteiller[3] and H. W. W. Spoon[4]

[1]*Astronomy Centre, Department of Physics and Astronomy, University of Sussex, Falmer, Brighton BN1 9QH, UK*
[2]*Virginia Polytechnic Institute & State University, Department of Physics, MC 0435, 910 Drillfield Drive, Blacksburg, VA 24061, USA*
[3]*Laboratoire AIM, CEA/DSM-CNRS-Universite Paris Diderot DAPNIA/Service d Astrophysique Baöt. 709, CEA-Saclay, F-91191 Gif-sur-Yvette Cedex, France*
[4]*Department of Astronomy and Center for Radiophysics and Space Research, Cornell University, Space Sciences Building, Ithaca, NY 14853-6801, USA*

Accepted 2013 October 2. Received 2013 October 2; in original form 2013 May 5

ABSTRACT

The mid-infrared (MIR) spectra observed with the *Spitzer* Infrared Spectrograph (IRS) provide a valuable data set for untangling the physical processes and conditions within galaxies. This paper presents the first attempt to blindly learn fundamental spectral components of MIR galaxy spectra, using non-negative matrix factorization (NMF). NMF is a recently developed multivariate technique shown to be successful in blind source separation problems. Unlike the more popular multivariate analysis technique, principal component analysis, NMF imposes the condition that weights and spectral components are non-negative. This more closely resembles the physical process of emission in the MIR, resulting in physically intuitive components. By applying NMF to galaxy spectra in the Cornell Atlas of Spitzer/IRS sources, we find similar components amongst different NMF sets. These similar components include two for active galactic nucleus (AGN) emission and one for star formation. The first AGN component is dominated by fine structure emission lines and hot dust, the second by broad silicate emission at 10 and 18 μm. The star formation component contains all the polycyclic aromatic hydrocarbon features and molecular hydrogen lines. Other components include rising continuums at longer wavelengths, indicative of colder grey-body dust emission. We show an NMF set with seven components can reconstruct the general spectral shape of a wide variety of objects, though struggle to fit the varying strength of emission lines. We also show that the seven components can be used to separate out different types of objects. We model this separation with Gaussian mixtures modelling and use the result to provide a classification tool. We also show that the NMF components can be used to separate out the emission from AGN and star formation regions and define a new star formation/AGN diagnostic which is consistent with all MIR diagnostics already in use but has the advantage that it can be applied to MIR spectra with low signal-to-noise ratio or with limited spectral range. The seven NMF components and code for classification are available at https://github.com/pdh21/NMF_software/.

Key words: galaxies: star formation – galaxies: statistics – infrared: galaxies.

1 INTRODUCTION

Spectra of the integrated mid-infrared (MIR) emission from galaxies contain a wealth of diagnostics that probe the origin of their MIR luminosity. For example, the main polycyclic aromatic hydrocarbons (PAHs) emission features found at 6.2, 7.7, 8.6, 11.3 and 12.7 μm are strong in objects where star formation (SF) activity contributes significantly to the MIR luminosity (Genzel et al. 1998; Laurent et al. 2000). The PAH features are either weak or absent for objects dominated by an active galactic nucleus (AGN) while emission lines with a high-ionization potential, for example the neon fine structure line [Ne V] 14.3 μm, tend to be strong in the presence of an AGN (Genzel et al. 1998; Sturm et al. 2000). Ratios of other fine structure lines such as [Ne III] 15.56 μm/[Ne II] 12.81 μm versus [S III] 33.48 μm/[Si II] 34.82 μm have been shown to diagnose power source (Dale et al. 2006) as has the shape of the underlying MIR dust continuum. (Brandl et al. 2006).

⋆ E-mail: p.d.hurley@sussex.ac.uk

Observations from the *Infrared Space Observatory* (Kessler et al. 1996) and the Infrared Spectrograph (IRS; Houck et al. 2004) on the *Spitzer Space Telescope* (Werner et al. 2004) allowed the MIR spectral features to be used as diagnostics of SF and AGN activity. Combinations of the PAH emission lines, high- to low-excitation MIR emission lines, silicate features and continuum measurements have been used as diagnostics for characterizing the power source behind ultraluminous infrared galaxies (ULIRGs; Genzel et al. 1998; Rigopoulou et al. 1999; Farrah et al. 2007, 2008, 2009; Spoon et al. 2007; Petric et al. 2010).

However, diagnostics based on specific emission and absorption lines only focus on small parts of the spectrum, disregarding the information contained in the rest of the MIR region. They can also be ambiguous. Dust and gas require ionizing radiation to emit in the MIR; the source of the radiation is not important. For example, hot OB stars or an accretion disc around a supermassive black hole can both produce the [O IV] 25.9 μm emission line, as well as shocks (e.g. Lutz et al. 1998). The line ratios of fine structure lines can also be affected by the geometry of the emitting region and the age of a starburst, while the metallicity can affect PAH emission strength (e.g. Thornley et al. 2000; Engelbracht et al. 2005; Madden et al. 2006; Wu et al. 2006; Farrah et al. 2007). As a result, different diagnostics can give conflicting estimates for the contribution from SF and/or AGN (e.g. Armus et al. 2007; Veilleux et al. 2009).

Separation of spectral features from continuum and the mixing of neighbouring spectral features can also be problematic. For example, measurement of the 9.7 μm silicate feature requires different methods depending on the strength of the 8.6 and 11.2 μm PAH emission lines (Spoon et al. 2007).

An alternative method for identifying the power source is to decompose the spectra with AGN and starburst spectral templates. These templates tend to be a spectrum from a specific object (e.g. M82) or a mean spectrum of a number of similar object types. Pope et al. (2008) use a combination of the M82 spectrum, average spectral template of starburst galaxies (Brandl et al. 2006) and a power law to decompose the IRS spectra of 13 high-redshift submillimetre galaxies. Valiante et al. (2009) fit IRS spectra across the range 5.5–6.85 μm with a combination of the M82 spectrum and a linear approximation for the AGN continuum. Alonso-Herrero et al. (2011) use the Brandl et al. (2006) starburst template and CLUMPY radiative transfer models for AGN to decompose the IRS spectra of 53 LIRGs into starburst and AGN components. Using average starburst templates is both simplistic and problematic. Prior theoretical prejudices drive the choice for what objects are used for the average templates, and they may be contaminated by AGN emission. The same is true for AGN average spectral templates.

With the public release of all low-resolution *Spitzer*/IRS spectra by the Cornell Atlas of Spitzer/IRS sources (CASSIS; Lebouteiller et al. 2011),[1] we are now in a better position to investigate the role played by SF and AGN with more sophisticated techniques. In this paper, we use a multivariate analysis technique to blindly learn the fundamental MIR spectral components, which we interpret as different physical environments within galaxies. Learning the MIR spectral shape of physical environments, allows the whole MIR wavelength range to be used as a diagnostic. The spectral components also provide an alternative to average spectral templates.

A subclass of multivariate analysis techniques include matrix factorization algorithms. The techniques are often associated with pattern recognition and blind source separation (Lee & Seung 2001). Algebraically, the algorithms approximate a data matrix by two simpler matrices: a weight matrix and component matrix. Common factorization techniques include singular value decomposition, principal component analysis (PCA) and independent component analysis (ICA). The different techniques use different assumptions to carry out the factorization, resulting in different weights and components. As multivariate data sets of spectra have become more prevalent, techniques such as PCA have been applied to astronomical problems. PCA has already been used for spectral classification of optical galaxies (e.g. Connolly et al. 1995; Bromley et al. 1998; Taghizadeh-Popp, Heinis & Szalay 2012). PCA has also been successfully applied to the IRS spectra of local ULIRGs (Wang et al. 2011; Hurley et al. 2012).

The weights and spectral templates derived with PCA can be both positive and negative. Spectral reconstruction involves both addition and cancellation of spectral features. As a result, the PCA templates are inherently difficult to interpret physically.

A relatively new matrix factorization technique, non-negative matrix factorization (NMF; Lee & Seung 1999) can be thought of as PCA but with non-negative constraints on weights and templates. The constraints make reconstruction a purely additive process which more closely resembles emission in the MIR. The first application of NMF to astronomy was carried out by Blanton & Roweis (2007) who adopted the Lee & Seung (2001) NMF algorithm and applied it to optical spectra and photometry. It has also been used as a blind source separation algorithm on the IRS spectra of galactic photodissociation regions (PDRs; Berné et al. 2007; Rosenberg et al. 2011).

This paper presents the first NMF analysis on MIR galaxy spectra. We use spectra from the recently released CASSIS (Lebouteiller et al. 2011). Our paper provides the first large-scale statistical analysis of the IRS spectra to date using the NMF algorithm. Section 2 describes the CASSIS data base and data reduction. In Section 3, we describe the suitability of matrix factorization to IRS spectra, and give details on the NMF algorithm. In Section 4, we present our results and in Section 5 our conclusions. We assume a spatially flat cosmology with $H_0 = 70\,\mathrm{km\,s^{-1}Mpc^{-1}}$, $\Omega = 1$ and $\Omega_\mathrm{m} = 0.3$.

2 THE DATA

2.1 CASSIS

We use spectra from the CASSIS (Lebouteiller et al. 2011). The atlas contains sources observed in low-resolution mode with the IRS (Houck et al. 2004) on board the *Spitzer Space Telescope* (Werner et al. 2004). IRS low-resolution mode observations were made using two low-resolution modules, ShortLow and LongLow (hereafter SL and LL, respectively), covering 5.2–14.5 and 14.0–38.0 μm, respectively. The modules also had a resolving power of $R \approx 60$–120 (≈75 per cent of the observations) and an aperture size of 3.7×57 arcsec for SL and 10.7×168 arcsec for LL. The observations in the CASSIS data base are first processed with the basic calibrated data (BCD) pipeline from the *Spitzer* Science pipeline (release S18.7.0.) and produce BCD frames. This removes electronic and optical artefacts. The BCD images are then processed using the CASSIS pipeline which carries out image cleaning, background subtraction and spectral extraction. The pipeline algorithm is both automatic and flexible enough to handle different observations, from barely detected sources to bright sources and from point-like to somewhat extended sources.

[1] The CASSIS is a product of the Infrared Science Center at Cornell University, supported by NASA and JPL.

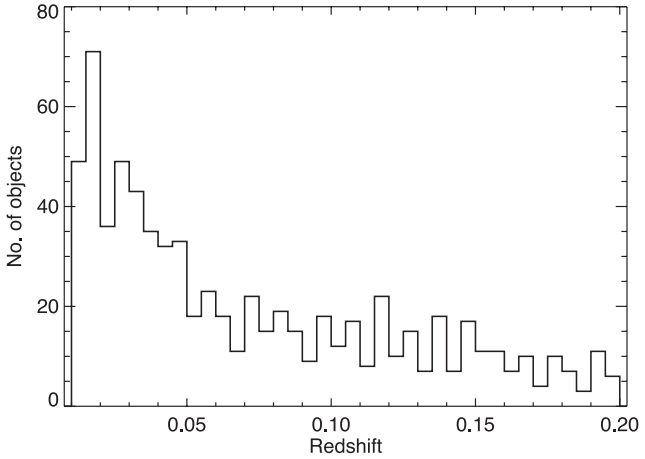

Figure 1. The redshift distribution for the sample selection we apply the NMF algorithm to.

2.2 Sample

The current version of CASSIS (version 4) contains 11 304 distinct sources. 2118 of those distinct sources have known spectroscopic redshifts taken from NASA/IPAC Extragalactic Database.[2] We make the additional redshift cut ($0.01 < z < 0.2$). The lower limit prevents contamination from Galactic and Local Group sources while the upper limit ensures that we sample approximately the same wavelength range for each object. The redshift cut gives us a sample size of 893. We note that the redshifts within CASSIS, have been collected heterogeneously, biasing our sample by the parent redshift surveys. Because objects in our sample are at low redshift and span many programmes, they likely span most or all infrared (IR) luminous object types in the local Universe. Therefore, while a small degree of bias is inevitable, we do not consider that it is significant enough to significantly affect our results. We also only use objects with both SL and LL data. This reduces our sample size down to 729 objects. The redshift distribution for the 729 objects can be seen in Fig. 1.

2.3 Stitching

Observations using data from both SL and LL spectral modules can suffer from mismatching due to telescope pointing inaccuracy or if a source is extended in SL and not in LL. The mismatching causes the spectra from one of the modules (normally the SL) to have lower flux calibration than the other. Correcting the mismatch is inherently difficult as the data from the overlap between the two modules can suffer from the '14 μm teardrop' (see IRS instrument handbook), [3] leaving a small gap at around 13–14 μm.

We correct for the mismatch using a simplified version of our NMF technique. For the first step, we generated two sets of templates, one using SL data and the other using LL data. The distribution in redshift causes the mismatch region to occur at different rest-frame wavelengths for different objects. This ensures that at least one template set covered the mismatch region for each object. We then fitted the template set to a region of width 7 μm, centred on the mismatch area. Wavelength points associated with PAH and neon emission lines were removed to prevent strong line strengths from distorting the fits. We carry out the fit for different scalings applied to the SL data. The scaling factor value that gives the lowest χ^2 is chosen as the scaling correction. Having stitched the spectra using both SL and LL template sets, we then generated our initial NMF sets for the entire spectral range. We then re-stitch the spectra with the new NMF set. Additional spectra used for analysis in this paper are also stitched with our final NMF_7 set, introduced in Section 4.

2.4 Normalization

The NMF analysis requires all spectra to be normalized to a standard value to prevent sources with higher flux, biasing the algorithm. We normalize all the spectra by the average flux across the rest-frame wavelength range of 7–20 μm. We choose this range as it is common to all sources with both SL and LL data.

3 MATRIX FACTORIZATION

Analysis of spectra from *Spitzer*'s IRS has tended to be done using diagnostics based on only a few of the specific features (e.g. Sajina et al. 2007; Pope et al. 2008; Alonso-Herrero et al. 2012). For example, Spoon et al. (2007) introduced a classification scheme based on the 6.2 μm PAH line and 10 μm silicate feature. Quantifying the contribution from SF and AGN has also been carried out using fine structure lines, for example the [O IV]/[Ne II] and [Ne V]/[Ne II] line ratios versus the 6.2 μm PAH equivalent width (EQW; e.g. Armus et al. 2007; Petric et al. 2010).

In essence, line diagnostic analyses are carrying out a crude compression by using only small parts of the spectrum to describe each object (e.g. the 6.2 μm feature). Matrix factorization techniques provide an alternative approach to compression by transforming data from wavelength space to one that better captures the variance in the data set. As a result, classification or quantification of properties such as SF is carried out considering a greater wavelength range.

Algebraically, matrix factorizations find a linear approximation to a data matrix $\mathbf{X}$ such that $\mathbf{X} \approx \mathbf{WH}$, or

$$X_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \quad (1)$$

Where, i is object index, μ is wavelength and a is component index. The matrix H can be thought of as a set of r components that represent latent structure explicit in the data set and W are a set of weighting coefficients. Each object in the data set can now be approximated by a linear combination of the derived components, H.

Different matrix factorization techniques use different assumptions to carry out the approximation. ICA assumes the derived components (H) are independent. PCA models the data set as a multivariate Gaussian distribution in wavelength space and finds the orthogonal components of the Gaussian. NMF assumes that the data, weights and components are all non-negative, but makes no assumption on the distribution of the data or correlation between derived components.

3.1 Matrix factorization of spectra

By applying linear matrix factorization techniques to the MIR spectra of galaxies, we are assuming that MIR spectra of galaxies, $F(\lambda)$, can be modelled as a linear combination of components. Ideally, the components would relate to physical regions, for example a star-forming region (T_{SF}), an AGNs torus (T_{AGN}), a molecular cloud

[2] http://nedwww.ipac.caltech.edu/
[3] http://irsa.ipac.caltech.edu/data/SPITZER/docs/irs/

($T_{\rm MC}$) or diffuse dust component ($T_{\rm C}$). A spectrum for a galaxy would then simply be

$$F(\lambda) = aT_{\rm SF}(\lambda) + bT_{\rm AGN}(\lambda) + cT_{\rm MC}(\lambda) + dT_{\rm C}(\lambda), \quad (2)$$

where, *a*, *b*, *c* and *d* are the relative weights for each component.

For the above model, ICA is not suitable as the components are unlikely to be independent, for example AGN and SF are believed to be triggered by similar mechanisms such as mergers (e.g. Sanders et al. 1988), and are likely to be connected through feedback processes (e.g. Farrah et al. 2012; Rovilos et al. 2012).

PCA has already been applied to the MIR spectra of ULIRGs (e.g. Wang et al. 2011; Hurley et al. 2012). Algebraically, PCA calculates the eigenvectors of the covariance matrix. For spectra, the principal components represent the principal variations from a mean spectral template. The components are therefore allowed to have features which are positive and negative, and are also allowed to have a negative weighting when fitting objects. The freedom to be both positive and negative does not mimic the process of emission in the MIR, resulting in components that are inherently difficult to interpret. By their nature, the principal components have a statistical rather than physical interpretation. Therefore, although PCA can successfully reduce dimensionality of spectra for classification from known objects, it is not suitable for our model.

The non-negative constraint of NMF more closely reflects the physical process of emission in the MIR, which does not suffer from the same problems of absorption as other spectral ranges. As a result the NMF generated templates are more physically intuitive.

NMF is therefore the most applicable matrix factorization routine for our linear interpretation of galaxy emission. However, the situation is complicated by dust extinction. This introduces a non-linearity to the problem since extinction is multiplicative and exponential.

$$F(\lambda) = (aT_{\rm SF}(\lambda) + bT_{\rm AGN}(\lambda) + cT_{\rm MC}(\lambda) + dT_{\rm C}(\lambda))e^{-f\tau(\lambda)}, \quad (3)$$

where *f* is the weight associated with extinction and $\tau(\lambda)$ can either be known or unknown.

We can take the model one step further by allowing extinction to vary across all four components:

$$F(\lambda) = aT_{\rm SF}(\lambda)e^{-f\tau(\lambda)} + bT_{\rm AGN}(\lambda)e^{-g\tau(\lambda)} + cT_{\rm MC}(\lambda)e^{-h\tau(\lambda)} + dT_{\rm C}(\lambda)e^{-i\tau(\lambda)}. \quad (4)$$

The weights for the extinction are *f*, *g*, *h* and *i*.

We have explored the suitability of non-linear kernel based matrix factorization algorithms (e.g. Zafeiriou & Petrou 2010; Pan, Lai & Chen 2011) and found they are not suited for the non-linear behaviour described in equations (3) and (4). We discuss why in Appendix A. Current algorithms therefore restrict us to describe MIR galaxy spectra as a set of linear components (e.g. equation 2) and NMF is the most appropriate matrix factorization technique.

The first application of NMF in astronomy was carried out by Blanton & Roweis (2007) who updated the popular NMF multiplicative algorithm from Lee & Seung (2001) to include uncertainties and for heterogeneous data sets (e.g. optical spectra and photometric observations of galaxies at different redshifts). They also restricted the space of possible spectra to those predicted from high-resolution stellar population synthesis models. We use the NMF algorithm from Blanton & Roweis (2007) to identify and learn the MIR sources that are common to galaxies in the CASSIS data base. Unlike Blanton & Roweis (2007), we do not use any models as a prior for shape of the components, we use the algorithm to blindly learn the shape of our components.

3.2 NMF algorithm

As with PCA, the goal of NMF is to minimize a cost function. The most widely used is the squared approximation error described in Lee & Seung (2001):

$$\chi^2 = \sum_{i\mu} \left(\mathrm{X}_{i\mu} - \sum_a \mathrm{W}_{ia}\mathrm{H}_{a\mu} \right)^2. \quad (5)$$

Minimizing equation (5) requires some sort of numerical technique to find local minima. Lee & Seung (2001) presented 'multiplicative update rules' for H and W. Upon each iteration, the rules are used to update H and W by a multiplicative factor whilst minimizing equation (5). The algorithm implemented in Blanton & Roweis (2007) altered the original multiplicative update algorithm of Lee & Seung (2001) for non-uniform uncertainties (σ). The cost function then becomes the weighted squared approximation error:

$$\chi^2 = \sum_{i\mu} \left(\frac{\mathrm{X}_{i\mu} - \sum_a \mathrm{W}_{ia}\mathrm{H}_{a\mu}}{\sigma_{i\mu}} \right)^2. \quad (6)$$

Blanton & Roweis (2007) showed the multiplicative update rules for H and W, which are as follows:

$$\mathrm{W}_{ia} \leftarrow \mathrm{W}_{ia} \left(\sum_\mu \frac{\mathrm{X}_{i\mu}\mathrm{H}_{a\mu}}{\sigma_{i\mu}^2} \right) \left(\sum_{m\mu} \frac{\mathrm{W}_{im}\mathrm{H}_{m\mu}\mathrm{H}_{a\mu}}{\sigma_{i\mu}^2} \right)^{-1} \quad (7)$$

$$\mathrm{H}_{a\mu} \leftarrow \mathrm{H}_{a\mu} \left(\sum_i \frac{\mathrm{W}_{ia}\mathrm{X}_{i\mu}}{\sigma_{i\mu}^2} \right) \left(\sum_{mi} \frac{\mathrm{W}_{ia}\mathrm{W}_{im}\mathrm{H}_{m\mu}}{\sigma_{i\mu}^2} \right)^{-1}. \quad (8)$$

The update rules in equations (7) and (8) are guaranteed to reduce the error; however, the cost function in equation (6) is not necessarily convex therefore the algorithm may get stuck in a local minimum. We run the algorithm five times with different initial starting positions to check that the solution is consistent.

Convergence can be evaluated by looking at the decrease in cost function across iterations and checking the solution has reached a minimum. In practise, we find 3000 iterations are enough for H and W to converge.

The number of components generated by NMF is a user input. Unlike PCA where the shape of the original components remain unchanged as more are added, the NMF components will not remain the same. We investigate the number of components required to constrain the data by generating 11 different NMF sets, containing from 3 up to 14 components. We define the following notation, NMF_y^x to describe the *x*th component from an NMF set containing *y* components.

3.3 Bayesian evidence

To determine the minimum number of components that are justified by the data, one should calculate the Bayesian evidence (*E*)

$$E \equiv \int L(\theta)\pi(\theta)\mathrm{d}\theta. \quad (9)$$

The evidence can be thought of as the average likelihood, $L(\theta)$, over all of the prior, $\pi(\theta)$, parameter space, $\mathrm{d}\theta$, of a given model and automatically implements Occam's razor, i.e. simpler models are preferred unless simplicity can be traded for greater explanatory power.

There are two ways in which one could calculate the Bayesian evidence for our setup. The first would be to calculate the evidence for the NMF algorithm, where the number of parameters is equal to the number of elements in both H and W. This approach would be the most appropriate if comparing the suitability of NMF to other matrix factorization techniques, the integral however becomes highly multidimensional making the calculation numerically challenging. Alternatively, if NMF is the most appropriate algorithm to our problem, then we can assume that the components are correct. The number of parameters is then equal to the number of elements in W, i.e. the number of components.

We choose the later approach as we have already chosen NMF as the most appropriate algorithm to our problem and are not comparing alternative procedures.

We calculate the evidence by using the nested sampling routine, MULTINEST (Feroz, Hobson & Bridges 2008) to re-fit the CASSIS sample with different NMF sets. MULTINEST is a Bayesian inference tool which calculates the evidence and produces posterior samples from distributions with (often an unknown number of) multiple modes and/or degeneracies between parameters. Nested sampling (Skilling 2004) is a Monte Carlo technique that randomly samples from the prior space, and zooms in on areas of higher likelihood during successive iterations.

We fit every galaxy with component sets NMF_3 to NMF_{14} and their respective repeats. For every repeat, we calculate the median evidence of the sample. The main uncertainty on our evidence values comes from the difference in NMF sets across repeats (i.e. the convergence on slightly different local minima by the NMF algorithm). To quantify the uncertainty on our evidence values, we calculate the mean and standard deviation evidence values from the five repeats, as a function of number of components.

4 RESULTS

4.1 Number of components

As discussed in the previous section, we would like to quantify how many components are required by the data. Fig. 2 shows the mean and standard deviation for the Bayesian evidence values from five repeats, as a function of number of components. The Bayesian evidence should start decreasing as the number of components exceeds the optimum number needed to constrain the data. We see no turnover, indicating there is not an obvious, optimum NMF set below 14 components. We note however a slight levelling off at seven components before increasing again beyond eight.

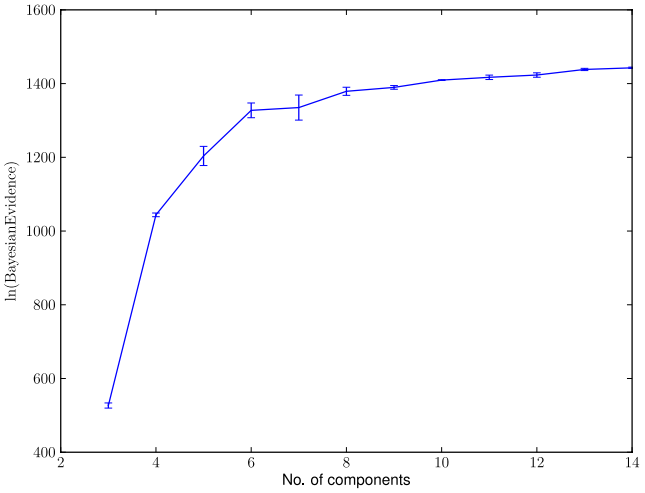

Figure 2. The Bayesian evidence as a function of number of components. For each NMF set, we run the algorithm five times and calculate the median evidence value of the entire galaxy sample. We plot the mean and standard deviation of the five repeats.

We have also looked at the ratio of evidence values between consecutive NMF sets. The ratio, referred to as the Bayes factor (K), is used as a measure for a Bayesian version of classical hypothesis testing. We use the Jeffreys scale to interpret K. A value of $K < 1$ indicates that the more complicated model is preferred, $K = 1$–3 as barely worth mentioning, $K = 3$–10 indicates substantial support for the simpler model, while $K = 10$–30 is strong, $K = 30$–100 is very strong and $K > 100$ is considered decisive. Using the Jeffreys scale, we find more than 14 components are needed to reconstruct spectra within the uncertainties. However, we note that K begins to level off after 6/7, indicating that although more complicated component sets are preferred, the gain in increasing the number of components is beginning to decrease.

Ideally, we would calculate the Bayesian evidence and Bayes factor beyond NMF_{14}. However, calculating evidence for highly multidimensional parameter spaces becomes computationally challenging. We have qualitatively examined NMF sets where number of components >14. As an example, in Fig. B1 we show the NMF components for NMF_{30}. Interpreting a many-component NMF set such as NMF_{30} becomes challenging as signatures begin to separate out into several components, whose physical interpretation is not clear.

We also note that the Bayesian evidence calculation could be influenced by two fundamental factors. The first is the use of uncertainties associated with IRS spectra, which have often been underestimated below the observed variation between individual nod positions on the IRS, as described in chapter 7 of the IRS Instrument Handbook.[4] As a result, our model selection may be too conservative. The other problem comes from the suitability of the NMF algorithm to the non-linear behaviour associated with extinction. We have carried out a simple simulation to show how extinction could be a factor in driving our linear methods to more templates than might be required by underlying physical conditions. Details can be found in Appendix B.

We have investigated how many components are needed in a quantitative manner. For the rest of this paper, we investigate the how many components are needed qualitatively, by examining some of the simpler NMF component sets, limiting our investigation to NMF_5–NMF_{10}.

4.2 Analysis of NMF_5 to NMF_{10}

Fig. 3 shows each spectral component for sets NMF_5–NMF_{10}. We have ordered the components so that similar components appear in the same order. We note that the ordering of components given by NMF is unimportant.

The NMF sets in Fig. 3 show that many of the components remain similar, despite an increase in the allowed number of components.

The first component contains a dust continuum which peaks at around 24 μm and contains emission from the sulphur line [S IV] at 10.51 μm, the 12.8, 15.6 and 24.3 μm neon lines and oxygen line [O IV] at 25.89 μm, all of which are associated with a hot ionized gas source. The continuum in the component from NMF_9 and NMF_{10}

[4] http://irsa.ipac.caltech.edu/data/SPITZER/docs/irs/

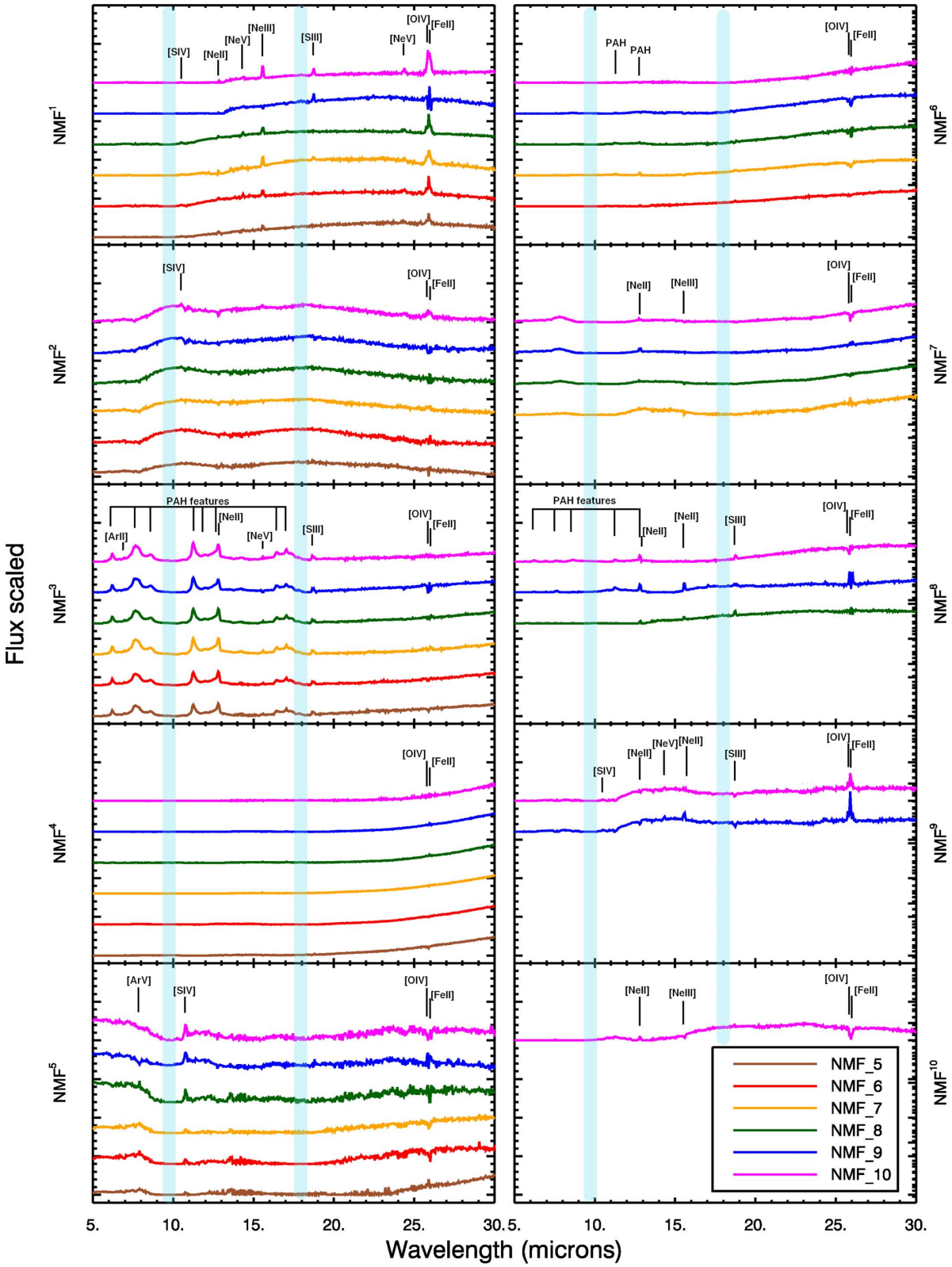

Figure 3. The derived NMF spectral components for sets NMF_5-NMF_{10}. Each NMF set is colour coded, with components ordered by similarity. For example, the five components of NMF_5 are the five brown spectra. Prominent spectral features in each component are also labelled and regions affected by broad silicate absorption and emission are highlighted in light blue.

varies from the others in that continuum does not start until 13 μm. This coincides with the appearance of the ninth and tenth components which show similar features. The hot dust continuum peaks at a wavelength similar to that of AGN tori, while the hot ionized gas emission lines have also typically been associated with AGN. The appearance of both in one component is consistent with the idea that they are correlated.

The second component shows silicate emission features at 10 and 18 μm due to stretching and bending of the Si–O and O–Si–O bonds, respectively. Silicate emission is typically associated with emission from very hot dust, found on the inner surface of AGN tori or narrow-line regions (Mason et al. 2009).

The third component captures the 6.2, 7.7, 8.6, 11.3, 12.7, 16.4 and 17.0 μm PAH features, and a cold dust slope at longer wavelength. There is also emission from argon line [Ar II] at 6.89 μm and sulphur line [S III] at 18.71 μm. Its shape is similar to the Brandl et al. (2006) average starburst template, based on 13 starburst galaxies. The ratio of the PAH features is very similar amongst component sets, but dust slope decreases with number of components. The reduction in dust slope for more complex NMF sets coincides with rising continuums seen in the fourth, sixth and seventh components.

The fifth component shows continuum emission up to 7 μm before dropping off at 10 μm. It also shows strong emission from the sulphur line [S IV] at 10.51 μm. The remainder of the spectrum is noisy and featureless.

The eighth, ninth and tenth components show similarities to the first component. They show varying amounts of emission from the neon lines, while the merged oxygen and iron lines appear as emission in the ninth component and absorption in the tenth. The variation of the first component in NMF_9 and NMF_{10} compared to the other NMF sets is a result of the introduction of the ninth and tenth components and occurs because the NMF algorithm is using the freedom of extra components to break down the first into subcomponents.

4.2.1 *Physical interpretation of the components*

The first two components both show features associated with hot dust and gas emission and are likely to be related to AGN emission. The unified model of AGNs predicts silicate emission from type 1 AGN and silicate absorption in type 2 AGN. More recently, the IRS spectra of type 2 quasi-stellar objects (QSOs) have shown silicate emission (Sturm et al. 2006). Schweitzer et al. (2008) have shown that the IRS spectra of 23 QSOs can be modelled with dusty narrow-line region models, while Mason et al. (2009) and Mor, Netzer & Elitzur (2009) showed that clumpy torus models could also provide silicate emission for both type 1 and type 2 AGN. The fact we see a relatively stable silicate emission component amongst different NMF sets would suggest that silicate emission is occurring in more than just type 1 AGN and is a fundamental spectral component.

The third component is the main SF component. It is dominated by PAH emission, often used as an indicator of SF (e.g. Roussel et al. 2001; Peeters, Spoon & Tielens 2004; Calzetti et al. 2005; Kennicutt et al. 2009), and predominantly comes from PDRs (Roussel et al. 2007; Peeters 2011). For simpler NMF sets, the component also contains a rising continuum at longer wavelengths due to colder dust emission ($T \approx 50$ K), also associated with SF (e.g. Calzetti et al. 2007). For the more complex NMF sets, the rising dust continuum is given its own component (e.g. the sixth and seventh). This indicates that although the colder dust and PAH emission both trace SF, they come from different regions and the NMF algorithm uses the additional freedom of extra components to separate the two. We note that the PAH emission is extremely stable amongst all NMF sets and we do not see significant PAH emission in any other component. Previous studies show the ratio of PAH features vary with metallically and radiation hardness (e.g. Smith et al. 2007), yet we have one component with PAH emission.

To investigate the stability and lack of variation in the PAH emission features, we have re-run the NMF algorithm on objects from our original sample which are dominated by the third component. Fig. 4 shows the components from NMF_4 to NMF_7 for our reduced sample. The NMF algorithm now finds two components with PAH emission. The first shows emission at 6.2, 7.7, 8.6, 11.3, 12.7, 16.4 and 17.0 μm, the second shows reduced emission for the 8.6, 11.3 and 12.7 μm PAH features and no emission at 16.4 and 17.0 μm, while at longer wavelengths there is a rising continuum. The two new PAH components show a resemblance to those found in an NMF analysis of IRS spectroimagery data for galactic PDRs (Berné et al. 2007). Their first component, interpreted as emission from deep within the PDR, showed broad emission at 6.2, 7.8 and 11.4 μm and a rising continuum. The second component contained emission from the 6.2, 7.6, 8.6, 11.3, 12.7 and 17.4 μm PAH features, and was shown to be more dominant in regions closer to the star.

By restricting the sample to objects dominated by SF, the NMF algorithm does not need to use components to separate out hotter dust from AGN, and uses the additional freedom to separate out the PAH emission. The PAH emission in our original third component is therefore capturing the average PAH emission from galaxies.

Components four, six and seven from Fig. 3, all contain rising continuums, though with varying slopes and are capturing dust emission at different temperatures. The fact we see numerous components with varying slopes suggests that the colder grey-body emission of dust varies considerably amongst galaxies. The seventh component also contains a bump at around 8 and 12 μm. The bumps help build up a silicate absorption feature at 10 μm, this component is therefore important for dusty galaxies.

To further investigate the components, we can begin to look at how they contribute to different types of spectra. In order to simplify the analysis and to provide a simple set of components, we restrict our components to those from NMF_7. Our choice of seven is more qualitative than quantitative, as we have already shown that a quantitative analysis requires more than 14 components. To validate our choice, we have studied the median, absolute residuals of NMF fits to the CASSIS sample with NMF_5 to NMF_9, shown in Fig. 5. The residuals are high for some of the emission lines, particularly the PAH features, because our components capture the average line emission. However, we note that by seven components, the residuals for the underlying continuum are down to 1σ and there is little advantage in using more complicated sets. By choosing seven, we believe we strike the balance between having enough simplicity to have a useful and physically intuitive NMF set of components, whilst being able to reconstruct the general spectral shape. The seven components are re-plotted in Fig. 6.

4.2.2 NMF_7 *fits to example galaxy spectra*

We now examine the NMF fits to spectra of different types of galaxies in order to show how contributions from components vary and that our NMF_7 set can capture the general shape of different types of spectra. Our example fits, along with the corresponding residuals (i.e. data fit) can be found in Figs C1 and C2. The first plot

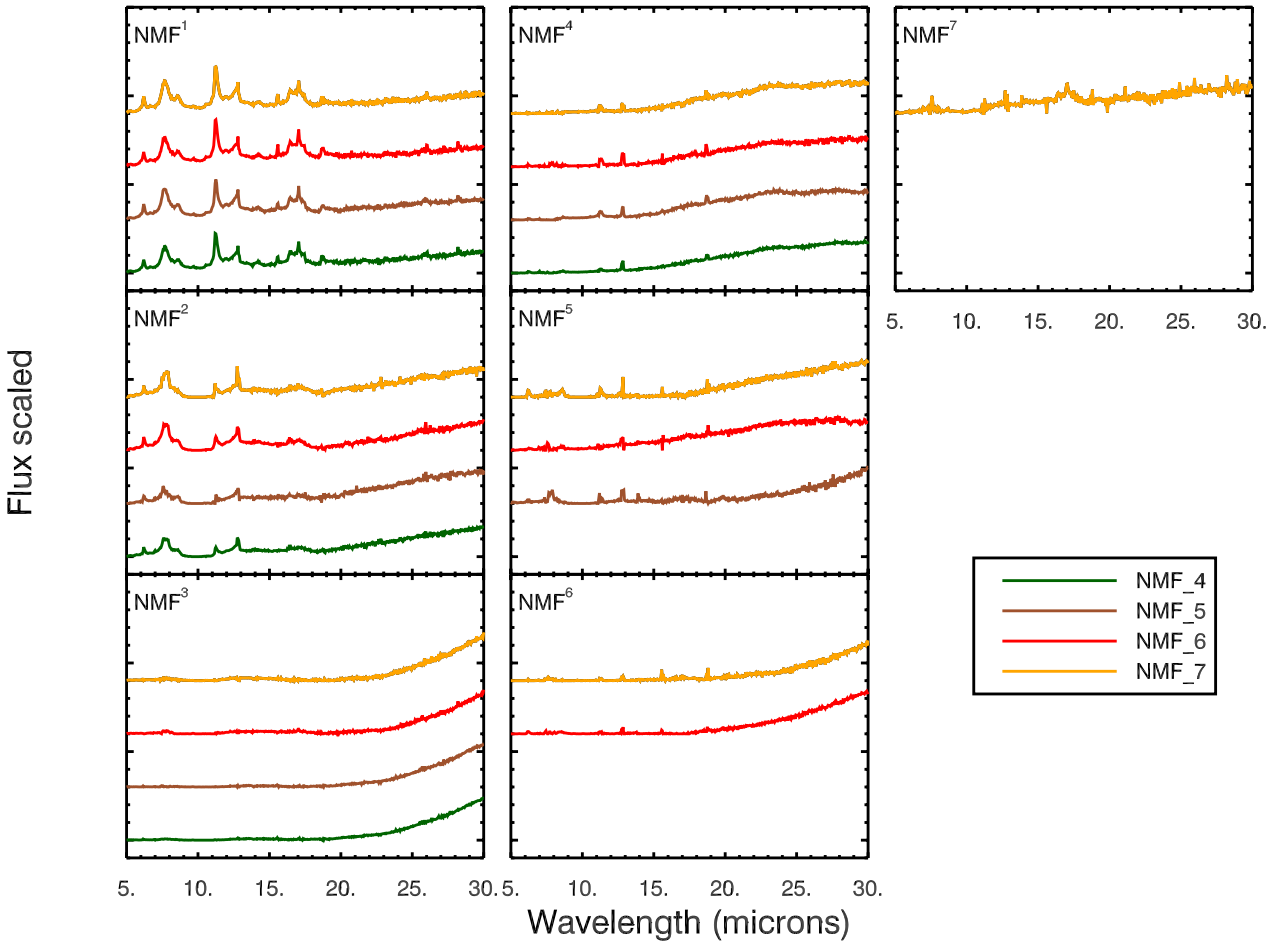

Figure 4. The derived NMF spectral components for NMF_4–NMF_7, using only objects dominated by the third PAH component seen in Fig. 3. Each NMF set is colour coded, with components ordered by similarity.

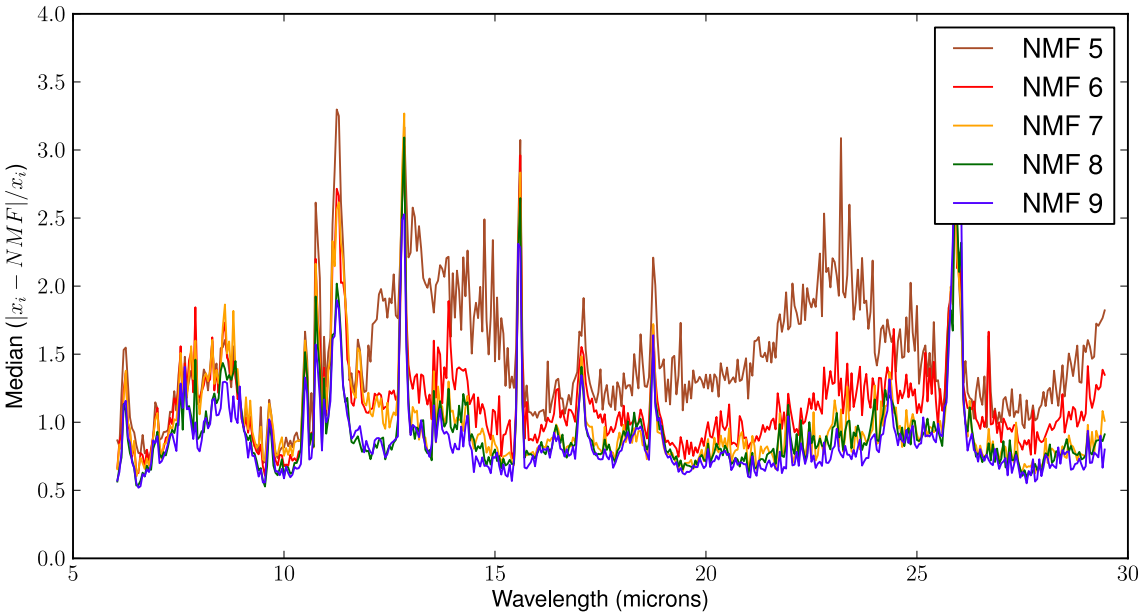

Figure 5. The median absolute residuals, normalized by σ, for NMF sets NMF_5–NMF_9. The residuals show that all NMF sets fail to capture the variance in many of the emission lines. However, for NMF sets NMF_7 and above, the residuals for the underlying continuum are down to 1σ.

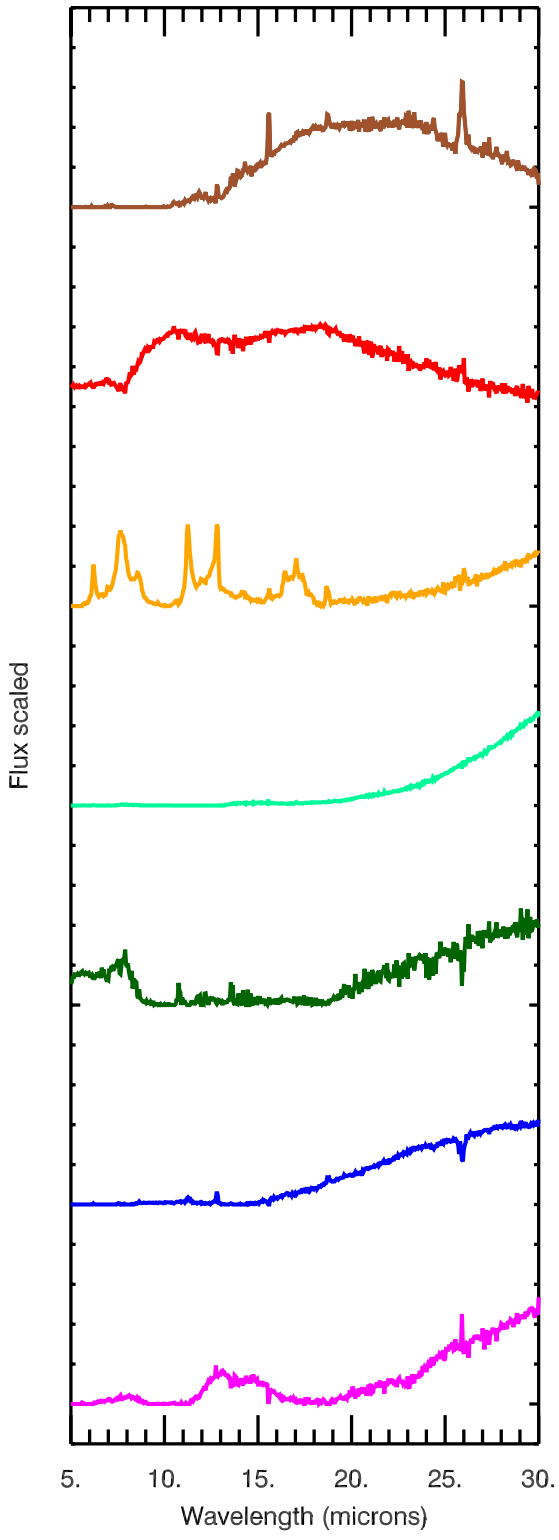

Figure 6. The seven components from NMF_7, corresponding to the yellow components in Fig. 3. The new colour coding is used to identify the different components in subsequent figures.

in Fig. C1 shows the NMF fit to the blue compact dwarf (BCD) KUG 1013+381, observed as part of the IRS Guaranteed Time Observation programme. BCDs tend to be small galaxies with low metallicity, that have undergone a recent burst of SF but have suppressed SF compared to typical starburst galaxies (Wu et al. 2006).

Our NMF fit shows that component one makes a significant contribution, suggesting there is some hot dust. Component four also makes a large contribution, indicating emission from colder dust. Components six and seven, both containing dust slopes at longer wavelengths, also contribute. There is very little emission from component two, which we believe is associated with the inner surface of an AGN and there is very little emission from the third 'PAH' component. The residual plot shows that the NMF_7 set can construct the underlying continuum; however, the [S IV], [Ne III] and [S III] emission lines are underestimated.

Our second NMF fit is to the ULIRG and type 1 Seyfert (Seyfert 1) galaxy, Markarian 231. Unlike, KUG 1013+381, the second 'silicate emission' component makes a contribution, and the other, warmer dust components such as six and seven contribute as much power to the longer wavelengths as the fourth component. There is very little contribution from the third component. Residuals show the fit is reasonable except beyond 25 μm, where there appears to be some instrumental artefact in the spectra.

The third fit is to PG 1211+143, also a Seyfert 1 galaxy. The second component dominates the emission of this object. The first, fifth and sixth components make comparable contributions. The residual plot shows that our NMF_7 set slightly overestimates emission from the [Ne III] and [O IV] lines.

The fit to the ULIRG and Seyfert 2 galaxy, Markarian 273, is dominated by emission from the fourth 'cold dust' component. Residuals show that the NMF components underestimate some of the emission lines, particularly the [Ne III] line. The continuum appears to be well reconstructed by the NMF components.

Our final two fits in Fig. C1 are to the starburst galaxies, NGC 3301 and NGC 3256. The third component contributes in the shorter wavelengths, while the colder dust components, four and six, contribute at longer wavelengths. The residuals show the components are capable of reconstructing the continuum, but fail to capture the emission lines accurately.

Four additional example fits are shown in Fig. C2. The first is to LINER, 3C270. The first, second and fifth components are the main contributors, while the residuals show the fit can reconstruct the continuum, but underestimate the 12.8 μm neon line. The submillimetre galaxy GN26 is over a short wavelength region and the spectrum is quite noisy. Our final two fits are to quasar PG 0804+761 and ULIRG IRAS 10378+1108. As with other type 1 AGN, the second component dominates emission. Our NMF_7 set fails to model the full width of the very broad silicate emission feature at 9.7 μm; however, the rest of the continuum is well reconstructed. Our NMF fit to the ULIRG IRAS 10378+1108 dominates the emission, while the residuals show the NMF set slightly overestimate the grey-body emission longwards of 27 μm.

In addition to galaxy spectra, we also fit our NMF_7 set to the average spectral templates from the IRS spectral ATLAS of galaxies (Hernán-Caballero & Hatziminaoglou 2011). Table C1 in Appendix C gives more details on the sources used for the ATLAS average templates. As can be seen in Fig. C3, the change in contributions for different types of object is consistent with those in Fig. C1. The continuum is well constructed for all average templates; however, the residuals show that the emission lines are not accurately reconstructed, especially for the average low-ionization nuclear emission-line region (LINER) template.

Overall, our fits show for Seyfert galaxies, the first and second component, along with the warmer dust components of five, six and seven are all important, though their contributions vary. For the starburst galaxies, the third and fourth components play a more important role. The residual plots show that our NMF set is capable

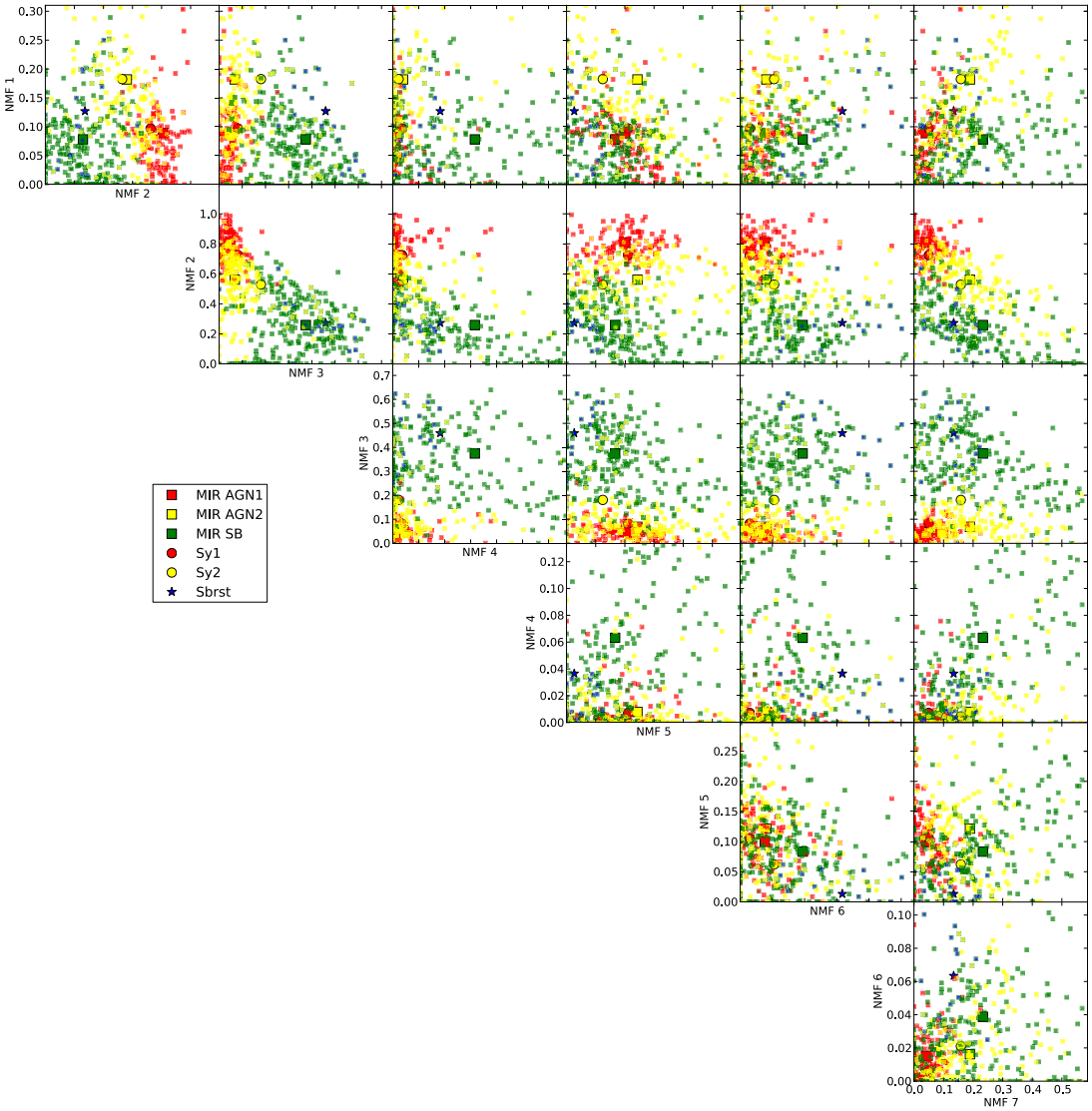

Figure 7. The distribution of objects/spectra from the ATLAS groups: MIR AGN1, MIR AGN2, MIR SB, Sbrst, Sy1 and Sy2 in our 7D space defined by the NMF_7 set. Symbols and colours for the different groups are described in the legend. The position of the average template for each group is marked by a larger symbol.

of reconstructing the continuum to a reasonable accuracy; however, some of the emission lines are not always fitted well. This is to be expected since, as we have previously shown, the components capture the 'average emission' of spectral lines. To accurately fit continuum and emission lines, our Bayesian evidence calculation has shown that we would need an NMF set with more than 14 components. The goal of this paper is to find a physically intuitive component set, which requires a balance between number of components and ability to reconstruct spectra. We believe Figs 5 and C1 show that our NMF_7 set fits this requirement.

To illustrate how the components contribute to a number objects, we can use the weightings provided by the NMF fits as multidimensional coordinates. Each galaxy is now a point in a 7D space we call NMF space. We use classifications from the IRS spectral ATLAS of galaxies (Hernán-Caballero & Hatziminaoglou 2011) to investigate what regions of NMF space are associated with different types of galaxies. The ATLAS collection contains spectra from a number of observing programmes. They provide optical classifications from the literature and three additional MIR classifications: MIR SB, MIR AGN1, MIR AGN2 based on the fractional contribution from a PDR component used during spectral decomposition. The AGN subgroups MIR AGN1 and MIR AGN2 are subsets of AGN, classified by whether spectra show silicate emission or silicate absorption. Fig. 7 shows how objects from the ATLAS groups: MIR AGN1, MIR AGN2, MIR SB, Sbrst, Sy1 and Sy2 are distributed in the 7D NMF space.

As can be seen in Fig. 7, the Seyfert 1 and MIR AGN1 objects all lie in a region with low contribution from NMF_7^1, high contribution from NMF_7^2 and very little contribution from NMF_7^3. The Seyfert 2 and MIR AGN2 objects are found in a region with a higher contribution in NMF_7^1, less or very little contribution from NMF_7^2 and very little contribution from NMF_7^3. Starburst like objects on the other hand require little contribution from either NMF_7^1 or NMF_7^2, and a high contribution from NMF_7^3.

We note that the components most influential in separating out the different objects are the components one, two and three. Less influential but still significant are the colder dust components NMF_7^4 and NMF_7^6. They contribute very little to objects classified as AGN,

Table 1. The percentage of ATLAS classified objects in each cluster for seven NMF templates. The first column indicates the cluster number. The second column shows the probability that a CASSIS object is in that cluster (i.e. how many objects can be found in it). The remaining columns contain the percentage of ATLAS classification in each cluster.

GMM and ATLAS classification for seven templates							
Cluster	Prob.	Sy1	Sy2	MIR AGN1	MIR AGN2	MIR SB	Sbrst
1	0.301	90.9	37.7	97.5	52.3	1.6	6.2
2	0.287	0.0	17.0	0.0	1.7	43.6	68.8
3	0.156	0.0	28.3	0.8	24.1	11.7	12.5
4	0.147	9.1	17.0	0.8	8.6	21.4	6.2
5	0.080	0.0	0.0	0.0	8.0	20.2	0.0
6	0.022	0.0	0.0	0.8	2.3	0.8	6.2
7	0.004	0.0	0.0	0.0	1.1	0.8	0.0
8	0.003	0.0	0.0	0.0	1.7	0.0	0.0

while the contribution for starbursts show a large variation. This fits in with our earlier interpretation that these two components represent obscured SF components which vary more than the PAH features seen in NMF_7^3. The remaining two components are the least significant. There is a slight difference in contribution between AGN 1 objects and the other two classes, while NMF_7^7 separates out type 1 and type 2 objects to a certain extent.

4.3 Gaussian mixtures modelling

We have shown NMF space is capable of separating out different types of objects. We now model how objects separate out in this multidimensional space by applying the parametric technique Gaussian mixtures modelling (GMM). GMM has already been successfully applied to the colour and redshift space of galaxies (Davoodi et al. 2006). GMM assumes the distribution of objects can be modelled by a series of clusters, each described by a multidimensional Gaussian. We use the GMM software from the Auton Lab[5] (Moore 1999) to model the distribution of the CASSIS sample in our 7D NMF space. The software uses the Expectation Maximization algorithm to learn the position and size of the clusters and uses the Akaike Information Criterion (AIC; Akaike 1974) to select how many are needed to describe the distribution of objects.

We find that eight clusters are required to adequately model the distribution. Each cluster describes a probability density function (PDF) for any position in NMF space. By using an object's position in NMF space, we can assign it to one of the eight clusters.[6] Table 1 shows how some of the ATLAS classified sources are distributed across the eight clusters, with clusters ordered by their normalization (i.e. how many objects are in that cluster). As can be seen in Table 1, the majority of objects are contained within the first five clusters. The normalizations associated with the remaining clusters (i.e. how many objects they capture) are also very small. We therefore use the first five clusters to define a classification scheme.

[5] http://www.autonlab.org

[6] Every position in NMF space has eight PDF values associated with it (one for each cluster). Using the highest probability density provides the optimal (maximum likelihood) classification. However, since the PDFs overlap, this will not provide the best classification for the population statistics. We therefore take the same approach as Davoodi et al. (2006) and randomly assign each galaxy to a cluster, with probability proportional to the PDF values at the galaxies position in NMF space.

The location in NMF space of the first five clusters can be seen in Fig. 8. Each cluster is represented by its 1σ contour. The CASSIS sample used for training the GMM is also plotted.

As can be seen in Fig. 8 and classifications in Table 1, cluster one captures nearly all the Seyfert 1 galaxies and some Seyfert 2 galaxies. Cluster two contains a significant number of objects previously classified as starbursts, while cluster three contains a large proportion of the remaining Seyfert 2 objects. The position of cluster four indicates this could be an intermediary group between typical type 1 and type 2 galaxies. The fifth cluster contains just over a fifth of those objects classified as starbursts in the MIR and no optically classified starbursts. Its position in NMF space also suggests that it captures those objects which are dusty starbursts.

We conclude that cluster one is related to Seyfert 1 galaxies, cluster two with starbursts, cluster three with Seyfert 2 galaxies and cluster four for galaxies showing signs of both Seyfert 1 and 2 (e.g. type 1.5). The fifth cluster captures those galaxies which are dusty and obscured. The clusters can be used as a classification scheme by taking any IRS galaxy spectrum, fitting with NMF_7 set and using the corresponding weights to identify what cluster the object is associated with.

We compare our classification scheme to the Spoon et al. (2007) diagram, which classified ULIRGs via the strength of their 9.7 μm silicate feature and 6.2 μm EQW. Fig. 9 shows 89 ULIRGs in the Spoon et al. (2007) diagram, colour coded by our classification. Seyfert 1 classified galaxies lie on the far left of the bottom horizontal branch, corresponding to a 1A and 1B Spoon classification, Seyfert 2 classified galaxies span the horizontal branch and 2B Spoon classification. The starburst classified objects are located in the far bottom right of the Spoon diagram, while dusty objects are spread out across the diagonal branch. Only three objects are classified as type 1.5 and they lie on the horizontal branch, in-between the Seyfert 1 and 2 classified galaxies.

Comparing the success rates of different classification schemes, without knowing the 'true' classification is always problematic; however, our classification scheme is consistent with the Spoon et al. (2007) interpretation of Fig. 9 in terms of the location of starbursts, AGN dominated objects and dusty objects. Unlike the Spoon diagram, our classification scheme can also distinguish between Seyfert 1 and 2 galaxies.

We have shown that our classification scheme is just as successful as the Spoon classification. However, our classification has three distinct advantages over Spoon et al. (2007). First, Spoon et al. (2007) only use the 9.7 μm silicate feature and 6.2 μm PAH EQW to separate out classes. By using the NMF components as a basis for our GMM based classification scheme, we make use of the whole MIR region to classify objects. This also enables us to classify objects where the 9.7 μm silicate feature and 6.2 μm PAH EQW are not available or difficult to measure. Secondly, our classification scheme is modelled on the number density of our CASSIS sample in NMF space. Since our sample contains a large variety of objects, any sample biases will have a small effect on the outcome of our classification scheme. The Spoon classes on the other hand, are chosen based on arbitrary cuts in the 9.7 μm silicate feature and 6.2 μm PAH EQW. Thirdly, because our clusters describe a PDF, we can give an indication of how likely a galaxy could be found in any one of the five clusters. For example, in Table 2 we show the probability of being in any of the five clusters for some famous objects.

We make our classification tool publicly available on the arXiv and at https://github.com/pdh21/NMF_software/.

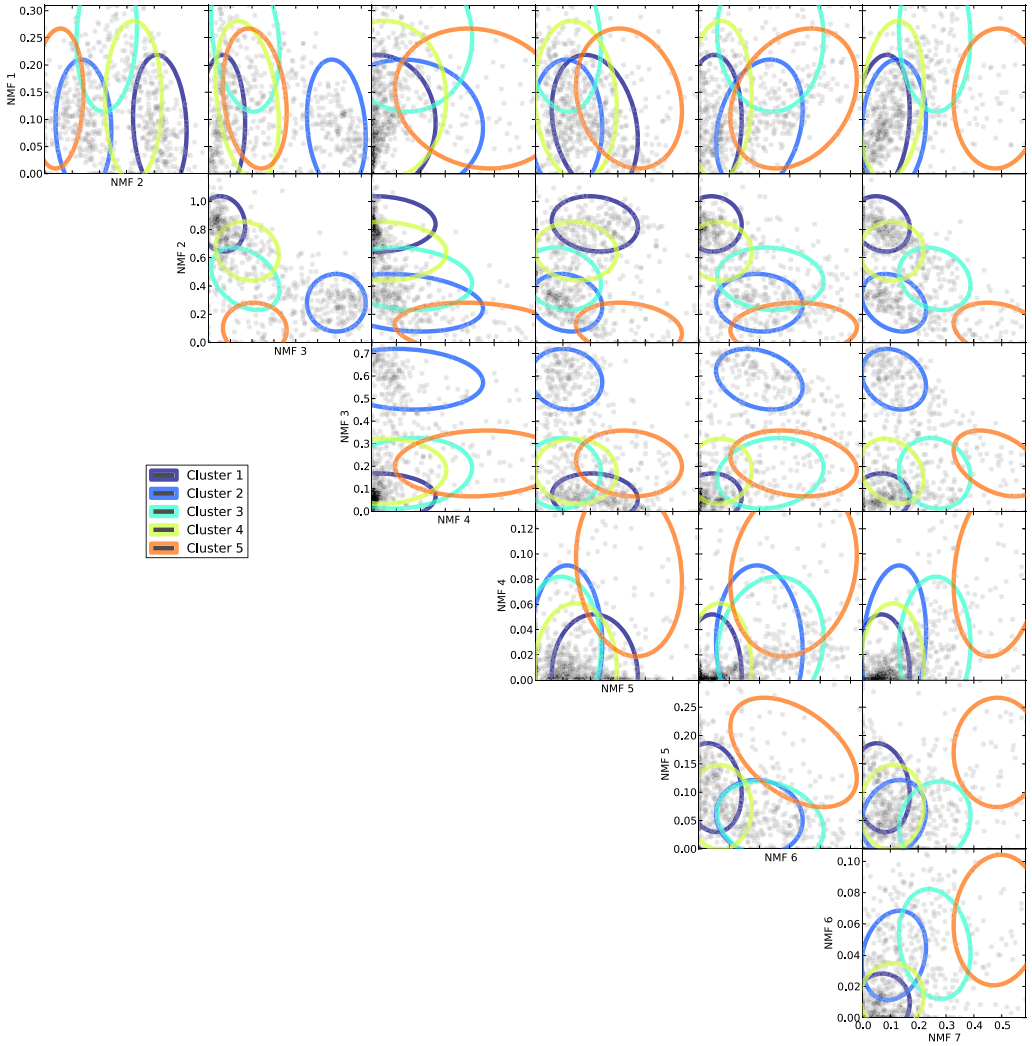

Figure 8. NMF space for seven templates. CASSIS objects used for NMF and GMM are also plotted. The ellipses represent the different clusters found through Gaussian mixtures modelling.

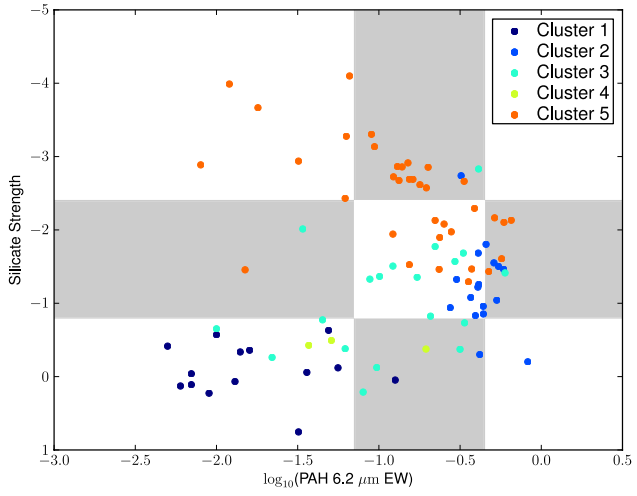

Figure 9. The Spoon et al. (2007) diagram showing silicate strength versus the 6.2 μm PAH EQW. The plot is separated into the different Spoon classes and objects are colour coded by our GMM classification.

Table 2. The approximate probability of being in one of the five clusters in our GMM based classification scheme.

Object	Cluster1 Sy1	Cluster2 Sbrst	Cluster3 Sy2	Cluster4 Sy1.5	Cluster5 Dusty SB
Arp 220	0.00	0.23	0.41	0.02	0.34
Mrk 231	0.32	0.00	0.34	0.32	0.02
PG 1211+143	0.92	0.00	0.00	0.08	0.00
IRAS 10565+2448	0.00	0.71	0.25	0.00	0.04
IRAS 10378+1109	0.00	0.01	0.06	0.00	0.93

4.4 SF–AGN contribution

We have shown that the NMF components are capable of distinguishing between the objects showing extreme SF or AGN activity. We now use them to introduce a diagnostic to quantify the contribution from SF and AGN. Unlike other diagnostics, ours employ the whole MIR spectrum to disentangle the SF versus AGN contributions, and it is not based on specific features for which we need to know information on their origin.

For AGN, NMF_7^1 and NMF_7^2 are the most important and bear the physical features we know to originate from AGN tori. We therefore

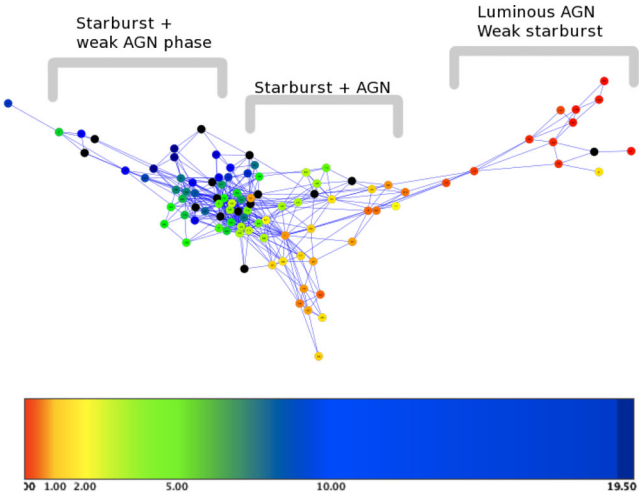

Figure 10. The network diagram along with interpretation from Farrah et al. (2009). Starbursts dominate the left-hand side of the network. As the AGN becomes more dominant, galaxies move to the right and finally on to one of the two branches. The nodes are colour coded by our NMF diagnostic. Nodes in black are where spectra are not available.

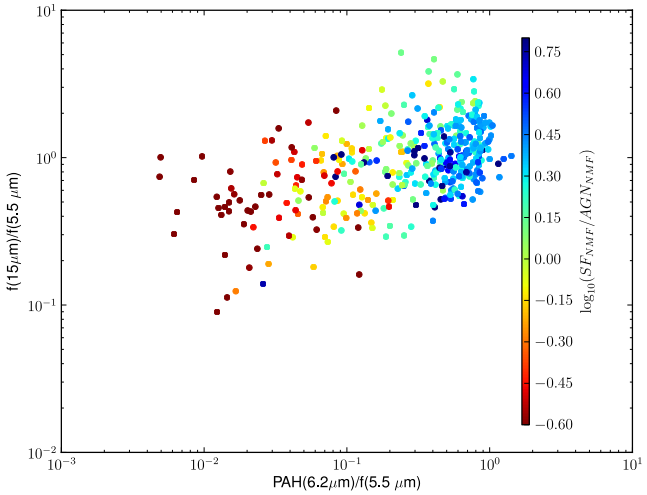

Figure 11. The ratio of 15 to 5 μm continuum flux, against the 6.2 μm PAH flux to 5 μm continuum flux, as seen in Armus et al. (2007). Points are colour coded by our NMF diagnostic.

adopt NMF_7^1 and NMF_7^2 as contribution from AGN. For SF, the third component is the most important; however, we argue that the fourth and fifth components are also required as they contain the colder dust associated with obscured SF. This is especially important for objects like Arp 220 which are known to be predominantly powered by SF but have less than average PAH emission compared to other submillimetre galaxies (Pope et al. 2008). We do not include NMF_7^6 and NMF_7^7 in our diagnostic. These components contribute to both AGN and starbursts and we have interpreted them as arbitrary dust components that are not specifically associated with SF or AGN activity. Our diagnostic is taken as the ratio of MIR luminosity from the following components:

$$\frac{\text{starformation}}{\text{AGN}} = \frac{L_{\mathrm{NMF3}} + L_{\mathrm{NMF4}} + L_{\mathrm{NMF6}}}{L_{\mathrm{NMF1}} + L_{\mathrm{NMF2}}}. \quad (10)$$

4.4.1 *Comparison to other MIR diagnostics*

We now show this diagnostic compared to other MIR diagnostic plots quantifying SF and AGN contribution.

Farrah et al. (2009) applied Bayesian inferencing and graph theory to a data set of 102 MIR spectra. By examining how position in the network was related to other parameters (e.g. IR luminosity, optical spectral type and black hole mass) they concluded that the network depicted the evolutionary scheme of ULIRGs, with different branches relating to starburst+AGN and luminous AGN.

We now investigate how our NMF_7 set relates to the same network by decomposing the Farrah et al. (2009) sample with our NMF components and colour coding the network by our NMF diagnostic. The connections are taken from Farrah et al. (2009) and we use the same CYTOSCAPE software[7] to produce the network. We note that our network is not identical to that in Farrah et al. (2009) due to the random seed starting position used by the spring-embedded algorithm in CYTOSCAPE. The two main branches seen in Farrah et al. (2009) are still seen in Fig. 10, with the lower and right-hand branches corresponding to the starburst+ AGN and luminous AGN branches, respectively. Each galaxy is colour coded by our new NMF diagnostic.

As can be seen in Fig. 10, our NMF diagnostic is consistent with the interpretation that SF occurs on the left-hand side of the network, with AGN activity increasing as we move to the right. The right-hand branch appears to be AGN dominated, as was concluded in Farrah et al. (2009).

Our second comparison is with the diagnostic diagram introduced by Laurent et al. (2000) and modified for *Spitzer* by Armus et al. (2007). The diagrams use the integrated continuum flux from 14 to 15 μm, the integrated continuum flux from 5.3 to 5.5 and the 6.2 μm PAH flux to indicate fractional contributions from AGN and starbursts. Fig. 11 shows the same diagnostic plot, plotted with objects from the CASSIS data base with measurements of the continuum and 6.2μm flux taken from the CASSIS data base. The points are colour coded by our NMF diagnostic.

Objects with a high NMF SF–AGN ratio are located in the top right while objects with a low NMF SF–AGN ratio lie in the bottom left. This is consistent with the simple linear mixing lines indicating AGN and SF fraction seen in Armus et al. (2007) and Petric et al. (2010).

Our third and fourth comparison is with diagnostic diagrams using emission lines. We plot all spectra in the CASSIS data base that have a known redshift and measurable emission line. Line measurements are made with the PAHFIT software (Smith et al. 2007). Fig. 12 shows the ratio of neon forbidden lines [Ne V] and [Ne II] against the PAH 6.2 μm EQW, colour coded by the NMF diagnostic. We indicate the fractional AGN and starburst contribution to the MIR luminosity from the [Ne V]/[Ne II] (vertical) and 6.2 μm PAH EQW (horizontal) assuming a simple linear mixing model. In each case, the 100, 50, 25 and 10 per cent levels are marked. The 100 per cent level is set by the average detected values for the [Ne V]/[Ne II] and PAH 6.2 μm EQW among AGN and starbursts, respectively, as discussed in Armus et al. (2007).

We see that our diagnostic is consistent with SF dominated objects being located in the bottom right of the plot, while objects with higher AGN contribution are located in the top left.

The third diagnostic diagram uses the [O IV] and [Ne II] ratio versus PAH 6.2 μm EQW. As in Fig. 12, we colour code the points by NMF diagnostic and indicate the fractional AGN and starburst

[7] Available from http://cytoscape.org/

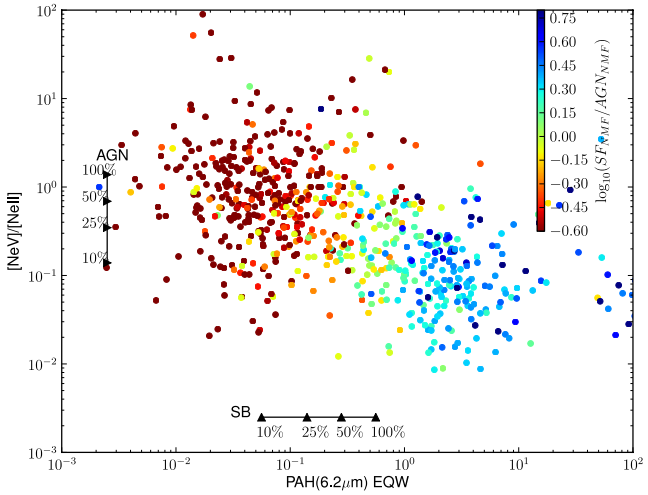

Figure 12. The [Ne v]/[Ne ii] ratio versus the PAH 6.2 μm EQW. The points are those objects in the CASSIS data base that have a redshift and an estimate for the three lines. The points are colour coded by our NMF diagnostic. We also show the 100, 50, 25 and 10 per cent AGN and starburst linear mixing contributions taken from Armus et al. (2007).

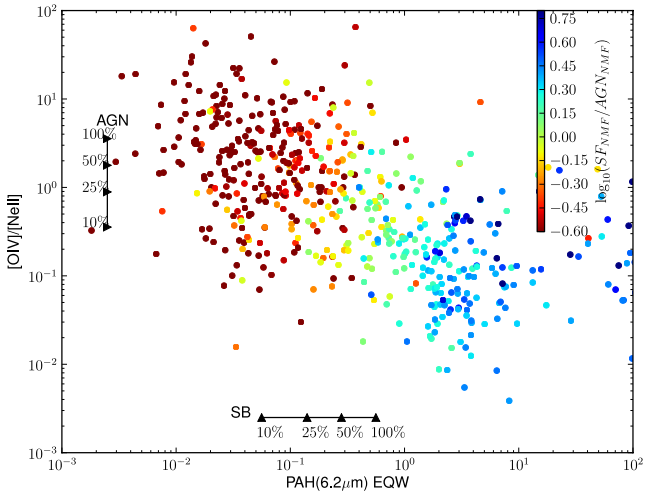

Figure 13. The [O iv]/[Ne ii] ratio versus the PAH 6.2 μm EQW. The points are those objects in the CASSIS data base that have a redshift and an estimate for the three lines. The points are colour coded by our NMF diagnostic. We also show the 100, 50, 25 and 10 per cent AGN and starburst linear mixing contributions taken from Armus et al. (2007).

contributions as discussed in Armus et al. (2007). Our plot can be seen in Fig. 13. AGN dominated objects lie the top left, SF dominated objects in the bottom right, which is consistent with the interpretation of Armus et al. (2007). Our final comparison is with Spoon et al. (2007) diagram, classifying ULIRGs via the strength of their 9.7 μm silicate feature and 6.2 μm EQW. Fig. 14 shows 89 ULIRGs in the Spoon et al. (2007) diagram, colour coded by our NMF diagnostic. Our NMF diagnostic suggests that AGN dominated objects are on the horizontal branch, while objects on the diagonal branch appear to have significant activity from SF and AGN. Objects dominated by SF lie at the extreme right of the two branches. Our diagnostic is consistent with the interpretation of Spoon et al. (2007).

We have shown that our diagnostic for determining the AGN/SF ratio is consistent with MIR diagnostic diagrams already in use.

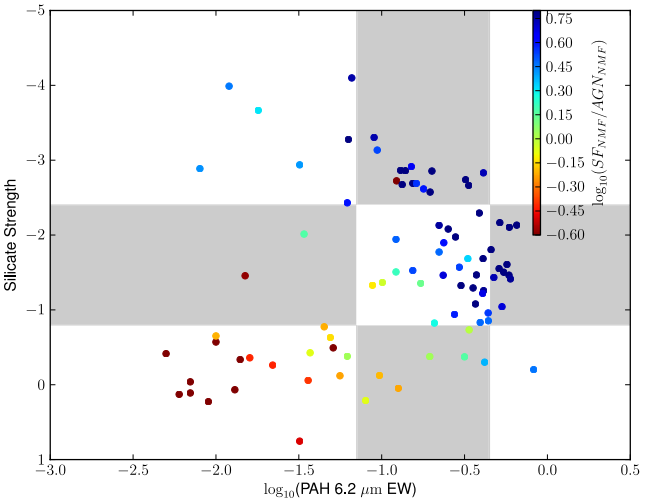

Figure 14. The Spoon et al. (2007) diagram showing silicate strength versus the 6.2 μm PAH EQW. The plot is separated into the different Spoon classes and objects are colour coded by the NMF diagnostic.

Our diagnostic however has the advantage that it uses a far greater wavelength range than current diagnostics and does not rely on specific line measurements. By using five of the seven components in NMF_7, our diagnostic is also flexible enough to account for the difference in spectra amongst SF or AGN dominated objects.

5 CONCLUSIONS

We have carried out the first empirical attempt at learning the fundamental MIR spectral components of galaxies via the multivariate analysis technique, NMF. We have chosen NMF as the most appropriate matrix factorization technique for our problem as the non-negative constraints required by the algorithm, more closely resemble the physical process of emission in the MIR than techniques used in previous studies (Wang et al. 2011; Hurley et al. 2012). The NMF algorithm has been applied to 729 galaxy spectra, taken from the CASSIS data base (Lebouteiller et al. 2011) with spectral redshifts ranging from ($0.01 < z < 0.2$).

We have investigated the number of components needed to accurately reconstruct spectra by evaluating the Bayesian evidence with the nested sampling routine, MULITNEST. The Bayes factor suggests that the number of components exceeds 14 but the gain in increasing the number of components decreases dramatically from seven components onwards. An NMF set with a large number of components may accurately reconstruct all spectra, but assigning physical interpretation to each component becomes difficult, limiting its practical utility.

We have therefore examined the simpler component sets NMF_5−NMF_{10}. We find that despite an increase in the allowed number of components, many of the components remain similar. For example, similar counterparts to components in NMF_5 can be found in NMF_6 and above, the sixth component in NMF_6 can be found in NMF_7 and above and so on. Finding similar components, despite an increase in flexibility, suggests these components are fundamental spectral components.

We find the components also have clear, physical interpretation. The first component contains the forbidden fine structure lines associated with narrow-line regions and AGN as well as a hot dust continuum also typical of AGN tori. The second common component shows silicate emission at 10 and 18 μm and is indicative of

the warm dust associated with both the inner wall of the AGN torus or narrow-line region clouds. The third component is a SF component, containing all of the PAH and molecular hydrogen emission lines, found near PDRs. As the number of components is increased, the colder dust slope is removed to the sixth and seventh components. We interpret this as the separation of unobscured star-forming component (or PDR) from an obscured star-forming component showing colder dust.

Re-running the NMF algorithm on objects dominated by SF, we show that the PAH emission begins to separate out into two components, which show similar features to the two different PDR components found in Berné et al. (2007).

We have shown that a simpler NMF set with seven components is capable of reproducing the general continuum shape for variety of extragalactic spectra seen in the MIR, though the components struggle with the variation in emission lines. By examining the contributions each component makes to well-known objects and previously classified samples, we find different types of objects lie in different regions of 'NMF space'.

Using GMM, we provide a classification scheme that uses all seven components to separate objects into five different clusters: a Seyfert 1 cluster, Seyfert 2 cluster, starburst cluster, dusty and obscured cluster and a Seyfert 1.5 cluster. Our classification outperforms the Spoon diagram in separating out Seyfert 1 and 2 like objects. Unlike the Spoon classification, ours use the whole MIR region, allowing objects without the 9.7 μm silicate feature and 6.2 μm EQW to be classified. Our GMM based classification can also provide an estimate of the probability of finding a particular galaxy in one of the five clusters.

We also use five of the components to create a SF/AGN diagnostic which performs well against current MIR diagnostic diagrams. Our NMF based diagnostic has the advantage of considering a greater wavelength range, and can therefore be used for objects where specific emission features have not been observed, or for where spectra are too noisy.

Our NMF components provide fundamental, physical components which are ideal for separating out different types of objects and investigating the power associated with AGN and SF. They are linked to the actual physical environments such as AGN and SF unlike templates based on specific objects (e.g. M82) or average templates based on a sample of galaxies. We believe our NMF set could be used to predict useful measures such as SF rate and AGN luminosity and will investigate this in a future paper. We also believe our NMF set is ideal for more galaxy evolution based investigations such as decomposing the MIR luminosity function into contribution from AGN and SF. Our NMF components and code for classification are made available at https://github.com/pdh21/NMF_software/.

ACKNOWLEDGEMENTS

We thank the referee for the useful comments, which have improved the paper. We acknowledge support from the Science and Technology Facilities Council (grant numbers ST/F006977/1 and ST/I000976/1). This work is based on observations made with the *Spitzer Space Telescope*, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA.

REFERENCES

Akaike H., 1974, IEEE Trans. Autom. Control, 19, 716
Alonso-Herrero A. et al., 2011, ApJ, 736, 82
Alonso-Herrero A., Pereira-Santaella M., Rieke G. H., Rigopoulou D., 2012, ApJ, 744, 2
Armus L. et al., 2007, ApJ, 656, 148
Berné O. et al., 2007, A&A, 469, 575
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Brandl B. R. et al., 2006, ApJ, 653, 1129
Bromley B. C., Press W. H., Lin H., Kirshner R. P., 1998, ApJ, 505, 25
Calzetti D. et al., 2005, ApJ, 633, 871
Calzetti D. et al., 2007, ApJ, 666, 870
Chiar J. E., Tielens A. G. G. M., 2006, ApJ, 637, 774
Connolly A. J., Szalay A. S., Bershady M. A., Kinney A. L., Calzetti D., 1995, AJ, 110, 1071
Dale D. A. et al., 2006, ApJ, 646, 161
Davoodi P. et al., 2006, AJ, 132, 1818
Engelbracht C. W., Gordon K. D., Rieke G. H., Werner M. W., Dale D. A., Latter W. B., 2005, ApJ, 628, L29
Farrah D. et al., 2007, ApJ, 667, 149
Farrah D. et al., 2008, ApJ, 677, 957
Farrah D. et al., 2009, ApJ, 700, 395
Farrah D. et al., 2012, ApJ, 745, 178
Feroz F., Hobson M. P., Bridges M., 2008, MNRAS, 384, 449
Genzel R. et al., 1998, ApJ, 498, 579
Hernán-Caballero A., Hatziminaoglou E., 2011, MNRAS, 414, 500
Houck J. R. et al., 2004, ApJS, 154, 18
Hurley P. D., Oliver S., Farrah D., Wang L., Efstathiou A., 2012, MNRAS, 424, 2069
Kennicutt R. C. et al., 2009, ApJ, 703, 1672
Kessler M. F. et al., 1996, A&A, 315, L27
Laurent O., Mirabel I. F., Charmandaris V., Gallais P., Madden S. C., Sauvage M., Vigroux L., Cesarsky C., 2000, A&A, 359, 887
Lebouteiller V., Barry D. J., Spoon H. W. W., Bernard-Salas J., Sloan G. C., Houck J. R., Weedman D. W., 2011, ApJS, 196, 8
Lee D., Seung S., 1999, Nature, 401, 788
Lee D., Seung S., 2001, Advances in Neural Information Processing Systems (NIPS). MIT Press, Cambridge, p. 556
Lutz D., Spoon H. W. W., Rigopoulou D., Moorwood A. F. M., Genzel R., 1998, ApJ, 505, L103
Madden S. C., Galliano F., Jones A. P., Sauvage M., 2006, A&A, 446, 877
Mason R. E., Levenson N. A., Shi Y., Packham C., Gorjian V., Cleary K., Rhee J., Werner M., 2009, ApJ, 693, L136
Moore A., 1999, in Kearns M., Cohn D., eds, Advances in Neural Information Processing Systems. Morgan Kaufman, San Francisco, CA, p. 543
Mor R., Netzer H., Elitzur M., 2009, ApJ, 705, 298
Pan B., Lai J., Chen W.-S., 2011, Pattern Recognition, 44, 2800
Peeters E., 2011, in Cernicharo J., Bachiller R., eds, Proc. IAU Symp. 280, The Molecular Universe. Cambridge Univ. Press, Cambridge, p. 149
Peeters E., Spoon H. W. W., Tielens A. G. G. M., 2004, ApJ, 613, 986
Petric A. O. et al., 2010, ApJ, 730, 28
Pope A. et al., 2008, ApJ, 675, 1171
Rigopoulou D., Spoon H. W. W., Genzel R., Lutz D., Moorwood A. F. M., Tran Q. D., 1999, AJ, 118, 2625
Rosenberg M. J. F., Berné O., Boersma C., Allamandola L. J., Tielens A. G. G. M., 2011, A&A, 532, 128
Roussel H., Sauvage M., Vigroux L., Bosma A., 2001, A&A, 372, 427
Roussel H. et al., 2007, ApJ, 669, 959
Rovilos E. et al., 2012, A&A, 546, 58
Sajina A., Yan L., Armus L., Choi P., Fadda D., Helou G., Spoon H., 2007, ApJ, 664, 713
Sanders D. B., Soifer B. T., Elias J. H., Madore B. F., Matthews K., Neugebauer G., Scoville N. Z., 1988, ApJ, 325, 74
Schwarz G., 1978, Ann. Stat., 6, 461
Schweitzer M. et al., 2008, ApJ, 679, 101
Skilling J., 2004, in AIP Conf. Proc. Vol. 735, 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 395
Smith J. D. T. et al., 2007, ApJ, 656, 770

Spoon H. W. W., Marshall J. A., Houck J. R., Elitzur M., Hao L., Armus L., Brandl B. R., Charmandaris V., 2007, ApJ, 654, L49
Sturm E., Lutz D., Tran D., Feuchtgruber H., Genzel R., Kunze D., Moorwood A. F. M., Thornley M. D., 2000, A&A, 358, 481
Sturm E. et al., 2006, ApJ, 653, L13
Taghizadeh-Popp M., Heinis S., Szalay A. S., 2012, ApJ, 755, 143
Thornley M. D., Schreiber N. M. F., Lutz D., Genzel R., Spoon H. W. W., Kunze D., Sternberg A., 2000, ApJ, 539, 641
Valiante E., Lutz D., Sturm E., Genzel R., Chapin E. L., 2009, ApJ, 701, 1814
Veilleux S. et al., 2009, ApJS, 182, 628
Wang L., Farrah D., Connolly B., Connolly N., LeBouteiller V., Oliver S., Spoon H., 2011, MNRAS, 411, 1809
Werner M. W. et al., 2004, ApJS, 154, 1
Wu Y., Charmandaris V., Hao L., Brandl B. R., Bernard-Salas J., Spoon H. W. W., Houck J. R., 2006, ApJ, 639, 157
Zafeiriou S., Petrou M., 2010, IEEE Trans. Image Process., 19, 1050

APPENDIX A: NON-LINEAR MATRIX FACTORIZATION TECHNIQUES

ICA, PCA and NMF are linear models and cannot efficiently model non-linearities such as dust extinction. Over the last decade, non-linear matrix factorization techniques have been developed to overcome certain non-linear situations. All of these non-linear based techniques use kernels to map data with non-linear structure into a kernel feature space, where the structure becomes linear. Techniques such as PCA or NMF can then be performed in the kernel feature space to recover the structure. These types of techniques are suited to problems where the non-linearity is of parametric form, e.g. points distributed along a circle. Dust extinction is exponential relationship unsuited to this type of technique (Pan, private communication).

APPENDIX B: NMF_{30} AND EXTINCTION SIMULATION

As described in Section 4.1, interpreting a many-component NMF set such as NMF_{30}, shown in Fig. B1, becomes challenging as signatures begin to separate out into several components, whose physical interpretation is not clear. One possible explanation is the non-linear process of extinction. To explore whether the extinction can cause problems with our NMF analysis, we simulate extinction via equation (3) described in Section 3.

Our simulation is divided into two parts. The first part assumes that galaxy spectra are a linear combination as described in equation (2), while the second assumes that equation (3) is valid. To simulate the spectra, we use NMF set NMF_5 and linearly combine them with weights randomly sampled from a distribution based on those found in the real sample. We do this 500 times to create 500 unique galaxy spectra.

The second part of our simulation involves adding extinction to the simulated spectra as described in equation (3) in Section 3. $\tau(\lambda)$ is defined by the Galactic Centre extinction law of Chiar & Tielens (2006).

We then carry out the NMF algorithm on both the unextincted and extincted spectra. We run the algorithm for NMF_5-NMF_{20} and use the simplified model selection measures: the AIC (Akaike 1974) and the Bayesian Information Criterion (BIC; Schwarz 1978), defined as follows:

$$\mathrm{AIC} \equiv -2\ln L_{\max} + 2k + \frac{2k(k+1)}{N-k-1} \tag{B1}$$

$$\mathrm{BIC} \equiv -2\ln L_{\max} + k\ln N, \tag{B2}$$

where $L_{\max}$ is the maximum likelihood solution, N is the number of data points and k is the number of parameters. A minimum value for the AIC and BIC correspond to the optimum model. Fig. B2 shows both the BIC and AIC for both sets of simulated spectra. As expected, the BIC and AIC indicate that the spectra without extinction can be adequately described by the NMF set with five components. For spectra with extinction, the BIC and AIC do not level off until $NMF_{15}-NMF_{20}$. This suggests that extinction could be a factor in driving our linear methods to more templates than might be required by underlying physical conditions.

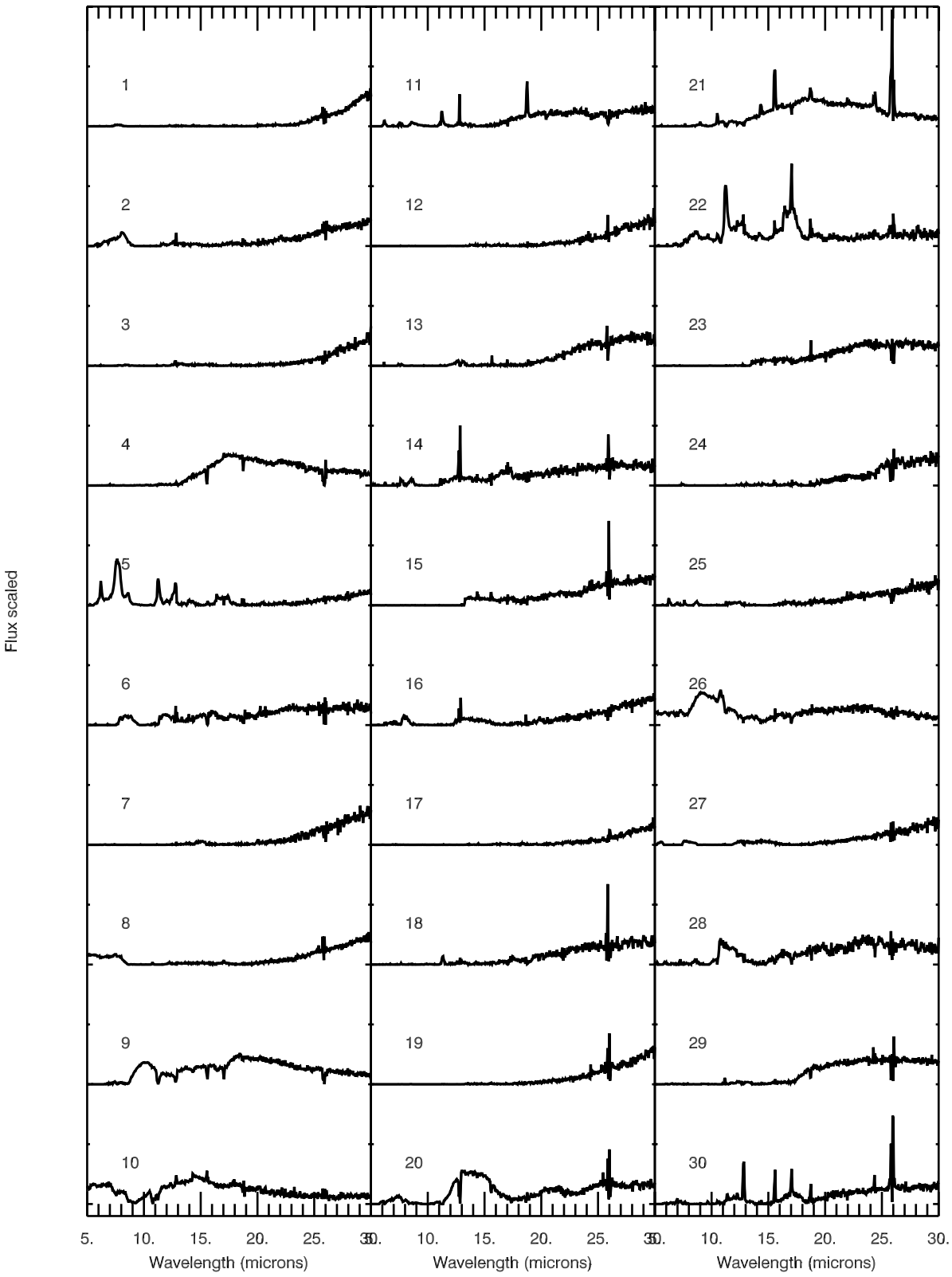

Figure B1. The 30 components of NMF set NMF_{30}.

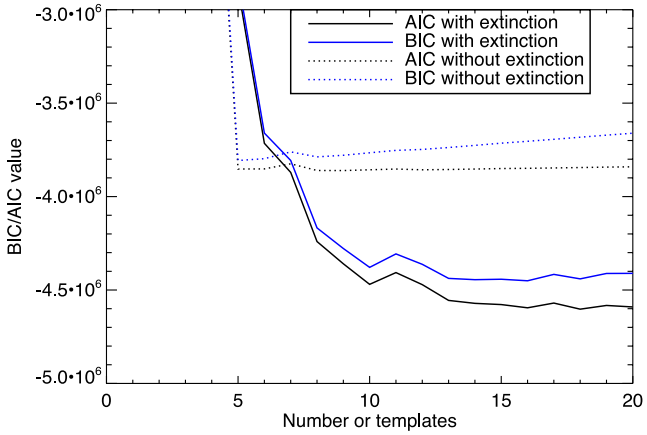

Figure B2. The AIC and BIC for the non-linear simulations. Both the BIC and AIC for spectra without extinction indicate five components as expected. The set with extinction requires around 15–20.

APPENDIX C: NMF$_7$ FITS TO GALAXY SPECTRA

As described in Section 4.2.2, we examine the NMF fits to spectra of different types of galaxies in order to show how contributions from components vary and that our *NMF*$_7$ set can capture the general shape of different types of spectra. Our example fits, along with the corresponding residuals (i.e. data fit) can be found in Figs C1, C2 and C3. Table C1 lists details of the ATLAS sample used in Hernán-Caballero & Hatziminaoglou (2011).

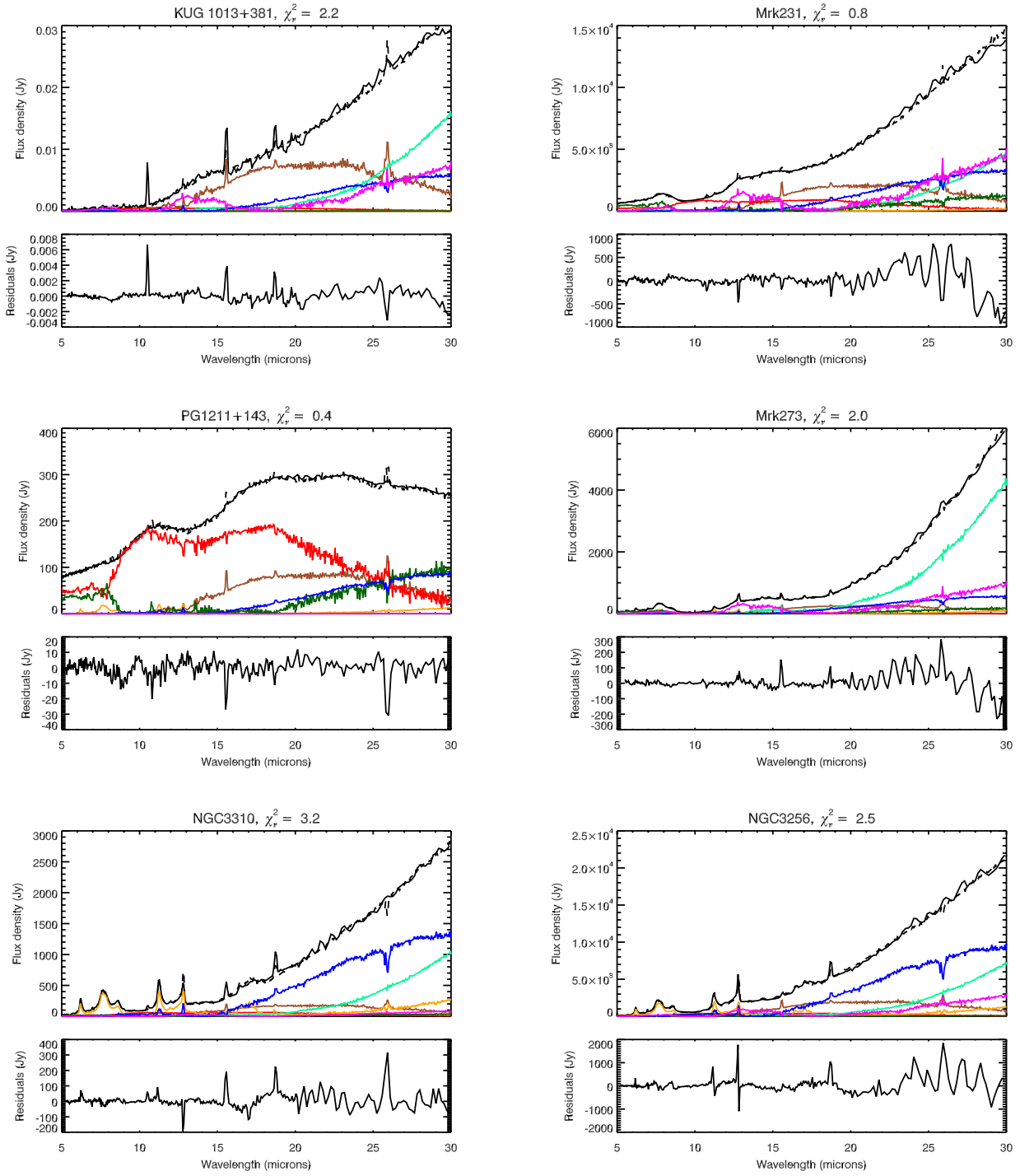

Figure C1. NMF_7 fits to the blue compact dwarf: KUG 1013+381, Seyfert 1 galaxies: Markarian 231 and PG 1211+143, Seyfert 2 galaxy: Markarian 273, and starburst galaxies: NGC 3310 and NGC 3256. Each spectrum is plotted as a black solid line and the NMF fit as black dashed line. The contribution from each component is also shown, with the same colour coding as in Fig. 6. The residuals (data fit) are plotted below each fit.

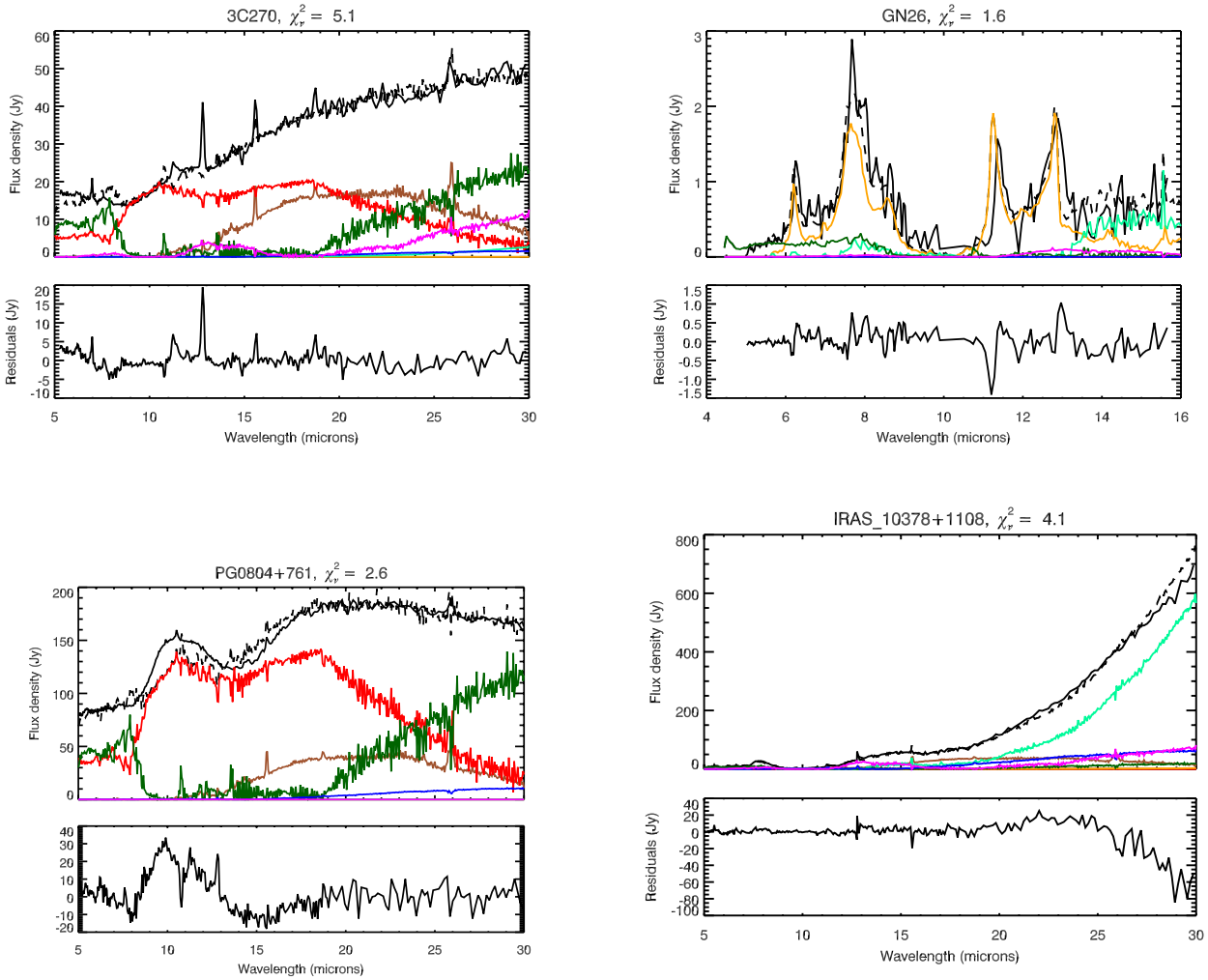

Figure C2. Four additional examples of NMF_7 fits. LINER: 3C270, submillimetre galaxy: SMG GN26, quasar: PG 0804+761 and ULIRG: IRAS 10378+1108. Each spectrum is plotted as a black solid line and the NMF fit as black dashed line. The contribution from each component is also shown, with the same colour coding as in Fig. 6.

Table C1. Classification of sources by Hernán-Caballero & Hatziminaoglou (2011).

Name	No. of sources	z_{min}	$\langle z \rangle$	z_{max}	λ_{min} (μm)	λ_{max} (μm)	Comments
Sy1	11	0.002	0.041	0.205	5.2	24.6	Seyfert 1 with $\nu L\nu(7\ \mu m) < 10^{44}$ erg s^{-1}
Sy1x	72	0.003	0.091	0.371	5.0	24.6	Intermediate Seyfert types (1.2, 1.5, 1.8, 1.9)
Sy2	53	0.003	0.045	1.140	5.2	24.6	Seyfert 2 with $\nu L\nu(7\ \mu m) < 10^{44}$ erg s^{-1}
LINER	16	0.001	0.034	0.322	5.2	24.6	LINER with $\nu L\nu(7\ \mu m) < 10^{44}$ erg s^{-1}
QSO	125	0.020	1.092	3.355	2.5	24.6	QSO1 and Seyfert 1 with $\nu L\nu(7\ \mu m) > 10^{44}$ erg s^{-1}
QSO2	65	0.031	1.062	3.700	3.6	24.6	QSO2 and Seyfert 2 with $\nu L\nu(7\ \mu m) > 10^{44}$ erg s^{-1}
Sbrst	16	0.001	0.091	1.316	5.2	24.6	Starburst or HII with $\nu L\nu(7\ \mu m) < 10^{44}$ erg s^{-1}
ULIRG	184	0.018	0.730	2.704	4.5	24.6	ULIRG (low- and high-redshift sources)
SMG	51	0.557	1.869	3.350	4.8	12.0	Submillimetre Galaxies
MIR_AGN1	119	0.002	0.455	2.190	4.0	24.6	MIR selected AGN with silicate emission
MIR_AGN2	160	0.002	0.549	2.470	4.5	24.6	MIR selected AGN with silicate absorption
MIR_SB	257	0.001	0.413	2.000	4.6	24.6	MIR selected starbursts

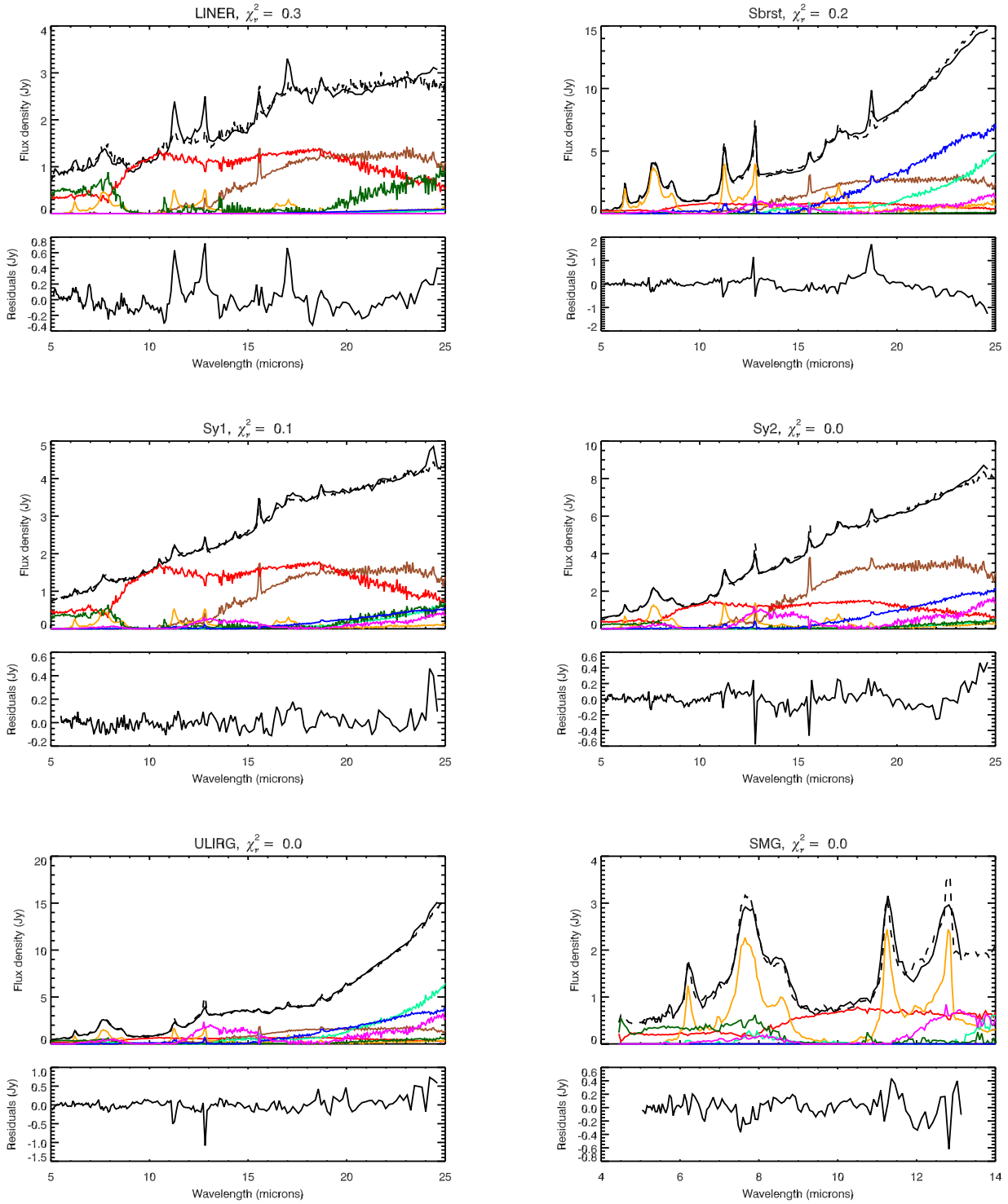

Figure C3. NMF$_7$ fits to the average templates from Hernán-Caballero & Hatziminaoglou (2011) (information on sample can be found in Table C1). Each spectrum is plotted as a black solid line and the NMF fit as black dashed line. The contribution from each component is also shown, with the same colour coding as in Fig. 6.

This paper has been typeset from a TEX/LATEX file prepared by the author.